QUT
Library

HIERARCHICAL MODELS FOR 2D PRESENCE/ABSENCE DATA HAVING AMBIGUOUS ZEROES: with a biogeographical case study on dingo behaviour

Samantha Jane Low Choy
BSc(Hons)(UQ)
Centre in Statistical Science and Industrial Mathematics
School of Mathematical Sciences

A THESIS SUBMITTED FOR THE
DEGREE OF DOCTOR OF PHILOSOPHY
OF THE QUEENSLAND UNIVERSITY OF TECHNOLOGY

submitted 6 December 1999
revised 5 March 2001

Gardens Point Stack
A20580975B
Hierarchical models for
2D presence/absence
data having ambiguous
zeroes : with a
biogeographical case

QUEENSLAND UNIVERSITY OF TECHNOLOGY

DOCTOR OF PHILOSOPHY THESIS EXAMINATION

CANDIDATE NAME	***Samantha Low Choy***
CENTRE/RESEARCH CONCENTRATION	***Statistical Science & Industrial Mathematics***
PRINCIPAL SUPERVISOR	***Prof Tony Pettitt***
ASSOCIATE SUPERVISOR(S)	***Dr Mervyn Thomas***
THESIS TITLE	***Bayesian Hierarchical Modelling of Two-Dimensional Lattice Data: A Biogeographical Case Study***

Under the requirements of PhD regulation 9.2, the above candidate was examined orally by the Faculty. The members of the panel set up for this examination recommend that the thesis be accepted by the University and forwarded to the appointed Committee for examination.

Name: AN PETTITT *Signature.*
Panel Chairperson *(Principal Supervisor)*

Name: VO ANH *Signature.*
Panel Member

QUT Verified Signatures

Name: Dawei HUANG *Signature.*
Panel Member

Name: KERRIE MENGERSEN *Signature.*
Panel Member

Under the requirements of PhD regulation 9.15, it is hereby certified that the thesis of the above-named candidate has been examined. I recommend on behalf of the Thesis Examination Committee that the thesis be accepted in fulfillment of the conditions for the award of the degree of Doctor of Philosophy.

Name DLS McElwain *Signature..* QUT Verified Signature *Date* 16/3/07
Chair of Examiners (External Thesis Examination Committee)

Keywords

2D lattice data, ambiguous zeroes, Autologistic distribution, 3-parameter autologistic, Bayesian inference, Bernoulli-Autologistic, binary Markov random fields, binary data, biogeography, bootstrapping, dingo behaviour, distribution maps, EM algorithm, environmental management, frequentist inference, Gibbs sampling, hierarchical model, Markov chain Monte Carlo, MCMC, Metropolis-Hastings, falsely inflated zeroes, Ising model, Normalization constant, population atlas, presence/absence data, spatial statistics, spatio-temporal data, statistical modelling, underlying spatio-temporal dependence.

Abstract

This dissertation is primarily an applied statistical modelling investigation, motivated by a case study comprising real data and real questions. Theoretical questions on modelling and computation of normalization constants arose from pursuit of these data analytic questions.

The essence of the thesis can be described as follows.

Consider binary data observed on a two-dimensional lattice. A common problem with such data is the ambiguity of zeroes recorded. These may represent zero response given some threshold (presence) or that the threshold has not been triggered (absence). Suppose that the researcher wishes to estimate the effects of covariates on the binary responses, whilst taking into account underlying spatial variation, which is itself of some interest. This situation arises in many contexts and the *dingo, cypress* and *toad* case studies described in the motivation chapter are examples of this.

Two main approaches to modelling and inference are investigated in this thesis.

The first is frequentist and based on generalized linear models, with spatial variation modelled by using a block structure or by smoothing the residuals spatially. The EM algorithm can be used to obtain point estimates, coupled with bootstrapping or asymptotic MLE estimates for standard errors.

The second approach is Bayesian and based on a three- or four-tier hierarchical model, comprising a logistic regression with covariates for the data layer, a binary Markov Random field (MRF) for the underlying spatial process, and suitable priors for parameters in these main models. The three-parameter autologistic model is a particular MRF of interest. Markov chain Monte Carlo (MCMC) methods comprising hybrid Metropolis/Gibbs samplers is suitable for computation in this situation. Model performance can be gauged by MCMC diagnostics.

Model choice can be assessed by incorporating another tier in the modelling hierarchy. This requires evaluation of a normalization constant, a notoriously difficult problem. Difficulty with estimating the normalization constant for the MRF can be overcome by using a path integral approach, although this is a highly computationally intensive method.

Different methods of estimating ratios of normalization constants (NCs) are investigated, including importance sampling Monte Carlo (ISMC), dependent Monte Carlo based on MCMC simulations (MCMC), and reverse logistic regression (RLR). I develop an idea present though not fully developed in the literature, and propose the Integrated mean canonical statistic (IMCS) method for estimating log NC ratios for binary MRFs. The IMCS method falls within the framework of the newly identified path sampling methods of Gelman & Meng (1998) and outperforms ISMC, MCMC and RLR. It also does not rely on simplifying assumptions, such as ignoring spatio-temporal dependence in the process. A thorough investigation is made of the application of IMCS to the three-parameter Autologistic model. This work introduces background computations required for the full implementation of the four-tier model in Chapter 7.

Two different extensions of the three-tier model to a four-tier version are investigated. The first extension incorporates temporal dependence in the underlying spatio-temporal process. The second extensions allows the successes and failures in the data layer to depend on time. The MCMC computational method is extended to incorporate the extra layer. A major contribution of the thesis is the development of a fully Bayesian approach to inference for these hierarchical models for the first time.

Contents

1 Preamble **1**
- 1.1 Introduction . . . 2
- 1.2 Scope of thesis . . . 2
- 1.3 Outline of thesis . . . 5
- 1.4 Contributions . . . 6

2 Motivation **9**
- 2.1 Introduction . . . 10
- 2.2 Zero-inflated data . . . 10
- 2.3 Motivating applications . . . 12
- 2.4 Overview of models for binary spatial data . . . 23
- 2.5 Discussion . . . 28

3 Frequentist hierarchical model **31**
- 3.1 Introduction . . . 33
- 3.2 Exploratory Data Analysis for *Dingo* case study . . . 33
- 3.3 Frequentist hierarchical modelling: transcript of published paper . . . 40
- 3.4 Discussion . . . 62

4 Binary Markov random fields **63**
- 4.1 Introduction . . . 65
- 4.2 Markov Random Field Models . . . 66
- 4.3 Anisotropic Autologistic model . . . 78
- 4.4 Simulation from MRFs . . . 95
- 4.5 Statistical inference for MRFs . . . 111
- 4.6 Discussion . . . 119

5 Bayesian Hierarchical Model **123**
- 5.1 Introduction . . . 124
- 5.2 Bayesian hierarchical model . . . 124
- 5.3 Inference . . . 128
- 5.4 Computations: MCMC Design . . . 130
- 5.5 Pilot simulation experiment . . . 132
- 5.6 Proposal experiment . . . 148
- 5.7 Bayesian model choice . . . 154
- 5.8 Discussion . . . 156

6 Normalization Constant **157**
6.1 Introduction 159
6.2 Simulation Strategies 163
6.3 Reverse Logistic Regression 171
6.4 Integrated Mean Canonical Statistic 176
6.5 Case study 197
6.6 Conclusions 215

7 Extended Hierarchical Bayesian Model **217**
7.1 Introduction 219
7.2 EXTENSION I 223
7.3 EXTENSION II 231
7.4 Results: overview 243
7.5 Results: EXTENSION I, Experiment 1 244
7.6 Results: EXTENSION I, Experiment 2 262
7.7 Results: EXTENSION II, Experiment 1 265
7.8 Discussion 271

8 Conclusions **273**
8.1 Summary of methodological results 274
8.2 Comparison of results for *Dingo* case study 275
8.3 Future directions 278

A Graphical representation of models **281**

B Bootstrapping estimates of s.e.'s **283**
B.1 Bootstrap estimates of sample errors 283
B.2 Alternative bootstrap sampling strategies 292

C MCMC diagnostics for *dingo* case study, Extension I **297**

D MCMC diagnostics Extension I, Experiment 2, *dingo* case study **299**

E MCMC diagnostics for *dingo* case study, Extension II **313**

List of Tables

1 Glossary of notation and abbreviations used often in the thesis. xvii

2.1 Design: dingo experiment. Tabulation of chemical types. 15

3.1 Frequency of dingo presence at a site with designated pair of chemicals, classified by chemical pair (row) and by day (column) 34
3.2 Frequency of dingo visits to locations at a site with designated chemical, classified by chemical (row) and by day (column). 35
3.3 Ranking of visits to chemical pair, by day. Superscripts indicate where visits exceed 2. 36
3.4 Ranking of visits to chemicals, by day 36
3.5 Equivalence between notation used in paper Pettitt & Low Choy (1999) and the rest of the thesis. 41
3.6 Number of chemically treated lures to which dingoes were attracted, classified by day and chemical 46
3.7 Chemical effect estimates for varying number of blocks over which probability of dingo presence is constant using ML 46
3.8 (a) Empirical probability transition matrix for chemical pair (A, B); and (b) Empirical marginal probabilities for y_1. 52
3.9 Bootstrap estimates and standard deviations for the EM analysis using an estimated marginal distribution for the first day. 53
3.10 Bootstrap estimates and standard deviation for the EM analysis using observed first day's data 53

4.1 Interpretation of 1st and 2nd order parameters in general pairwise interaction lattice models. 77

5.1 Choice of parameters in the prior distribution of dingo presence. 133
5.2 Spatio-temporal positions selected for monitoring dingo presence. 134
5.3 Pilot run: Statistics from posterior density of q_k. 138
5.4 Pilot run: Statistics from posterior density at selected positions of dingo presence $z_{s_j r_j}$ and summed over neighbouring positions. 141
5.5 Posterior mean of q_k for different starting values in dingo presence $\{z_{sr}^{(0)}\}$ . . 146
5.6 Percentage acceptance rates of q_k for different starting values in dingo presence $\{z_{sr}^{(0)}\}$ 146
5.7 IACT: Minimum, median, and maximum over q_k for different starting values in dingo presence. 147

5.8 Median IACT over selected dingo positions for different starting values in dingo presence. 147
5.9 Maximum IACT over selected dingo positions for different starting values in dingo presence $\{z_{sr}^{(0)}\}$ with different $\theta_{(m)}$. 147
5.10 Proposal distributions investigated for updates of q_k 148
5.11 Half-width parameters of uniform proposal distributions and corresponding standard deviations of Guassian proposals investigated for updates of q_k. . . 148
5.12 Half-width parameters of uniform proposal distributions and corresponding standard deviations of Guassian proposals investigated for updates of q_k. . 151

6.1 Comparison of Rectangular, Trapezoidal and Simpson's rule for integration of the mean canonical statistic in evaluating ratios of Normalization constants. 190
6.2 Triplets of low, middle and high conditional probabilities of presence (assuming positive dependence parameters) at a single site are given for various levels of prevalence (θ_0) and horizontal (θ_1) and vertical (θ_2) dependence. . 200
6.3 Results from estimation of log NC ratios using various methods, namely, Integrated Mean Canonical Statistic (IMCS), Importance Sampling Monte Carlo and Markov Chain Monte Carlo. 202
6.4 Simulation study results for Reverse Logistic Regression 205
6.5 Comparison of Reverse Logistic Regression and Integrated Mean Canonical Statistic. Base model $\theta_A = (-1.7, 0, 0)$ 214

7.1 Extension I, Experiment 1: Descriptive statistics of posterior distribution MCMC simulations of natural statistics for presence/absence $V(x)$. 260
7.2 Extension I, Experiment 1: MCMC Convergence Diagnostics for presence/absence natural statistics $V(x)$ 260
7.3 Extension I, Experiment 1: Descriptive statistics of posterior distribution MCMC simulations of the effects of explanatory variables q_k. 261
7.4 Extension I, Experiment 1: MCMC Convergence Diagnostics for effects of explanatory variables q_k 261
7.5 Extension I, Experiment 1: Cross correlations between each q_k chain 261
7.6 Extension II, Experiment 1: Design of Θ space. 265
7.7 Extension II, Experiment 1: Summary of Convergence diagnostics 266
7.8 Extension II: Categories of situations affecting the distribution of y_{ist} given y, z. 267
7.9 Summary of posterior distribution for q_k 267
7.10 Descriptive statistics for the simulated posterior distribution of each component of $V(x)$ 270
7.11 Descriptive statistics for the simulated posterior distribution of θ 270

8.1 Summary of chemical attractiveness, over extreme conditional and unconditional models and Frequentist models of Chapter 3 and the Bayesian models of Chapter 7. 276

B.1 Empirical distribution functions used to construct bootstrap samples. . . . 286
B.2 Number of EM iterations required to ensure that the given % change was achieved for each chemical effectiveness parameter $\{q_k; k = 1, \ldots, 6\}$. 288

B.3 Descriptive statistics of bootstrapped estimates of chemical effectiveness for smoothed and blocked estimates of the probability of dingo presence, and for various block sizes. 290
B.4 90% Interval estimates of bootstrapped estimates of chemical effectiveness for smoothed and blocked estimates of the probability of dingo presence, and for various block sizes. 291
B.5 Descriptive statistics of bootstrapped estimates of chemical effectiveness for smoothed and blocked estimates of the probability of dingo presence, for various block sizes. 294
B.6 Descriptive statistics of bootstrapped estimates of chemical effectiveness for smoothed and blocked estimates of the probability of dingo presence, for various block sizes. 295

C.1 Extension I, Experiment 1: Distributional choices for MCMC construction for Dingo experiment 297
C.2 Distributional choices for MCMC construction for experiment 2 298

D.1 Extension I, Experiment 2: MCMC descriptive statistics of posterior distribution simulations of the natural statistics for presence/absence $V(x)$. . . . 301
D.2 Extension I, Experiment 2: MCMC Convergence Diagnostics for presence/absence natural statistics $V(x)$ 301
D.3 Extension I, Experiment 2: MCMC descriptive statistics of posterior distribution simulations of the effects of explanatory variables q_k. 312
D.4 Extension I, Experiment 2: MCMC Convergence Diagnostics for effects of explanatory variables q_k 312

E.1 Extension II, Experiment 1: Descriptive statistics for posterior distribution of α, β 314
E.2 Extension II, Experiment 1: Convergence diagnostics for α, β 314
E.3 Extension II, Experiment 1: Cross correlations between each α_k, β chain . 318
E.4 Extension II, Experiment 1: MCMC descriptive statistics of posterior distribution simulations of the effects of explanatory variables q_k. 325
E.5 Extension II, Experiment 1: MCMC Convergence Diagnostics for effects of explanatory variables q_k 325
E.6 Extension II, Experiment 1: MCMC diagnostics of posterior distribution simulations of the natural statistics for presence/absence $V(z)$. 325
E.7 Extension II, Experiment 1: MCMC Convergence Diagnostics for presence/absence natural statistics $V(x)$ 326

List of Figures

2.1 Design: arrangement within a block 14
2.2 Differing opinions on spatial behaviour expected of dingos. 17
2.3 Relationship between ground-truthed, radar and true images, and associated parameters as used in Denham & Mengersen, 1999. 19
2.4 Alternative conceptual relationship between ground-truthed, radar and true image variables and other parameters in the *Cypress* case study. 22
2.5 Conceptual relationship between observed and true distribution maps, given proxies for research activity and a spatial parameter in the *Toad* case study. 22

3.1 Observations: positions where a dingo visited at least one of the locations so was definitely present. 34
3.2 Total dingo visits, to each site for each day in the experiment. 38
3.3 Total dingo visits or presences at each site. 39
3.4 Illustration of two blocks from the transect design. 44
3.5 The image of dingo presence/absence at the 135 sites over the 7 days. . . . 47

4.1 Geometric order of neighbourhood. 68
4.2 Notation for indexing sites in neighbourhood to site i. 68

5.1 Relationship between variables in initial hierarchical model for *Dingo* case study, with fixed θ. 126
5.2 Pilot run: Posterior distribution of q_1. 135
5.3 Pilot run: Posterior distribution of q_k, k=1,...,6. 136
5.4 Pilot run: batch means of q_k. Batch size 100. 137
5.5 Pilot run: batch variances of q_k. Batch size 100. 138
5.6 Pilot run: Starplots showing comparison of parameter estimates $\hat{q}_k$ between runs with different parameters in prior for dingo presence $\theta_{(m)}$. 140
5.7 Pilot run: posterior probability of a dingo being present. 142
5.8 Pilot run: posterior probability of a dingo being present 143
5.9 Reference to features of positions on the lattice monitored for convergence of $z_{sr}^{(t)}$. 143
5.10 Pilot run: Starplots showing comparison of parameter estimates $\hat{z}_{s_j r_j}$, between runs with different parameters in the prior for dingo presence $\theta_{(m)}$. . 144
5.11 Pilot run: Starplot showing comparison of parameter estimates $\hat{z}_{s_j r_j}$, averaged over neighbouring positions, between runs with different parameters in the prior for dingo presence $\theta_{(m)}$. 145
5.12 IACT for q_k: minimum, median and maximum IACT by $\theta_{(m)}$ and $R(q_k' \mid q_k)$ for half-widths in the range 0.01 to 0.10. 149

5.13 %Acceptance rates for q_k: minimum, median and maximum by $\theta_{(m)}$ and $R(q'_k \mid q_k)$ for half-widths in the range 0.01 to 0.10. 150
5.14 IACT for q_k: minimum, median and maximum IACT by $\theta_{(m)}$ and $R(q'_k \mid q_k)$ for half-widths in the range 0.05 to 0.50. 152
5.15 %Acceptance rates for q_k: minimum, median and maximum by $\theta_{(m)}$ and $R(q'_k \mid q_k)$ for half-widths in the range 0.05 to 0.50. 153

6.1 Variance of quadrature estimates of log NCR assuming equal variance at evaluation points. 188
6.2 The relationship between a pair of parameter values θ_A and θ_B which differ in 2 dimensions. 192
6.3 Log NC ratios compared to base model B obtained via Importance Sampling MC *vs* MCMC Ratio. Baseline model is B, $m = 1$ 208
6.4 Model diagnostics for fitting linear relationship between ISMC and MCMC Ratio Estimates. Baseline model is B, $m = 1$ 209
6.5 Comparison of log NC ratios obtained via MCMC Ratio and Importance Sampling Monte Carlo (ISMC) with different levels of dependence $\theta_{A,\text{tot}}$ annotated on the plot. 210
6.6 Sum and difference plot for comparison of ISMC and MCMC estimates of log NC with baseline model B, $m = 1$. 211
6.7 Model diagnostics for fitting linear relationship between ISMC and MCMC Ratio Estimates. Baseline model is C, $m = 17$ 212
6.8 Comparison of log NC ratios obtained via MCMC Ratio and IMCS with different levels of dependence θ_{tot} annotated on the plot. 213

7.1 Relationship between variables in extension I to hierarchical model for *Dingo* case study, with random θ. 221
7.2 Relationship between variables in extension II of hierarchical model for *Dingo* case study, with random θ. 222
7.3 Extension I, Experiment 1: Univariate posterior distribution of each component of θ. 246
7.4 Extension I, Experiment 1: Bivariate posterior distribution of (θ_1, θ_2), two components of θ. 247
7.5 Extension I, Experiment 1: Bivariate posterior distribution of (θ_0, θ_1), two components of θ. 248
7.6 Extension I, Experiment 1: Bivariate posterior distribution of (θ_0, θ_2), two components of θ. 249
7.7 Extension I, Experiment 1: Three-dimensional histogram of the joint marginal posterior distribution of the three θ components 250
7.8 Extension I, Experiment 1: Plot of joint posterior distribution of θ for each combination of spatial parameter θ_1 and temporal parameter θ_2. 251
7.9 Extension I, Experiment 1: Plot of joint posterior distribution of θ for each combination of prevalence parameter θ_0 and temporal parameter θ_2. 252
7.10 Extension I, Experiment 1: Plot of joint posterior distribution of θ for each combination of prevalence parameter θ_0 and spatial parameter θ_1. 253
7.11 Extension I, Experiment 1: Marginal proposal distribution of each discrete $\theta_{(m)}$. 254

7.12 Extension I, Experiment 1: MCMC Trace of the posterior distribution of natural statistics of presence/absence $V(x)$ 256
7.13 Extension I, Experiment 1: MCMC Trace of the posterior distribution of effects of explanatory variables q_1, q_2, q_3. 257
7.14 Extension I, Experiment 1: MCMC Trace of the posterior distribution of effects of explanatory variables q_4, q_5, q_6. 258

D.1 Extension I, Experiment 2: Posterior distribution of θ values compared to distribution of proposed θ^* values 300
D.2 Extension I, Experiment 2: Posterior distribution and MCMC time series of θ components 301
D.3 Extension I, Experiment 2: Joint posterior distribution of (θ_1, θ_2), components of θ. 302
D.4 Extension I, Experiment 2: Posterior distribution of θ_0 holding (θ_1, θ_2) constant.303
D.5 Extension I, Experiment 2: Joint posterior distribution of (θ_0, θ_1), components of θ. 304
D.6 Extension I, Experiment 2: Posterior distribution of θ_2 holding (θ_0, θ_1) constant. 305
D.7 Extension I, Experiment 2: Joint posterior distribution of (θ_0, θ_2), components of θ. 306
D.8 Extension I, Experiment 2: Posterior distribution of θ_1 holding (θ_0, θ_2) constant307
D.9 Extension I, Experiment 2: Distribution of prevalence parameter θ_0, compared to overall dependence as measured by $\theta_1 + \theta_2$. 308
D.10 Extension I, Experiment 2: MCMC Trace of the posterior distribution of natural statistics of presence/absence $V(x)$ 309
D.11 Extension I, Experiment 2: MCMC Trace of the posterior distribution of effects of explanatory variables q_1, q_2, q_3 310
D.12 Extension I, Experiment 2: MCMC Trace of the posterior distribution of effects of explanatory variables q_4, q_5, q_6 311

E.1 Extension II, Experiment 1: MCMC Trace of the posterior distribution of parameters in linear predictor $\alpha_1 - -\alpha_4$ 315
E.2 Extension II, Experiment 1: MCMC Trace of the posterior distribution of parameters in linear predictor $\alpha_5, \alpha_6, \beta$ 316
E.3 Extension II, Experiment 1: MCMC Autocorrelation function for simulated values in posterior distribution of parameters in linear predictor α_k, β . . . 317
E.4 Extension II, Experiment 1: CUSUM and hairiness diagnostic applied to simulation time series for parameter α_5 318
E.5 Extension II, Experiment 1: MCMC Trace of the posterior distribution of dingo presence canonical statistics $V_0(z)$ and $V_1(z)$ 319
E.6 Extension II, Experiment 1: CUSUM and hairiness diagnostic applied to simulation time series for parameter $V_1(z)$ 320
E.7 Extension II, Experiment 1: Acceptance probability accumulated over simulation time for statistics $V_0(z)$ and $V_1(z)$ 321
E.8 Extension II, Experiment 1: Frequency of accepted and proposed model indexes corresponding to $\theta_{(m)}$. 322
E.9 Extension II, Experiment 1: Frequency of accepted and proposed θ components (θ_0, θ_1) corresponding to $\theta_{(m)}$. 323

E.10 Extension II, Experiment 1: Trace of accepted and proposed model indices m corresponding to $\theta_{(m)}$ over simulation time. 324

List of Abbreviations

Table 1: Glossary of notation and abbreviations used often in the thesis.

Notation	Description
A	index for spatio-temporal process model, with parameter θ_A
B	index for spatio-temporal process model, with parameter θ_B
$c(\theta)$	generic normalization constant for distribution p having parameter θ
C	clique as defined in Section 4.2.2
C_H	specific heat related to the second derivative of the log NC
$\mathrm{E}[\cdot]$	Expectation
F	Free energy (thermodynamics)
G	number of grey levels in an image, or categories in a categorical z
H	magnetic field strength in statistical physics version of Ising model
$h(z,\theta)$	unnormalized probability distribution function
i	index of observation unit, relating to location in *dingo* case study
J	interaction term in statistical physics version of Ising model
k	index of factor (treatment) for *dingo* case study
k_B	Boltzmann's constant $= 1.38 \times 10^{-16}$ erg/deg
L	denotes maximum index for i with $i \in \mathcal{L} = \{1,\ldots,L\}$
L	sampling window $L \subset \mathcal{L}$ as used in Section 4.2.2
M	magnetization is related to the marginal total of Y
n_1	number of rows on the lattice
n_2	number of columns on the lattice
$p(\cdot\|\cdot)$	generic probability distribution function
$P(x,x*)$	transition probability density function (kernel) for transiting from old parameter value x to new value $x*$ for MCMC sampler in Section 4.4
q	expected probability of success
r	index for time as in y_{vsr}
R	maximum index for time as in $y_{vsr}, v = 1,\ldots,V; s = 1,\ldots,S; r = 1,\ldots,R$
$r(x*\|x)$	In Metropolis-Hastings version of MCMC, this is the proposal distribution for proposing a new parameter value $x*$ given the old value x, as discussed in Section 4.4

continued on next page

Table 1: (continued from previous page)

Notation	Description
S	summation shorthand $S(kzy) = \sum_{ist} I[\tau_{is} = k, Z_{st} = z, y_{ist} = y]$
s	index of observation unit, relating to spatial location
T	limit of s, with $s = 1, \ldots, S$.
t	index of simulation iteration, relating to parameter, *e.g.* $z_{sr}^{(t)}$
T	length of simulation and limit of t, with $t = 1, \ldots, T$.
U	internal energy (derivative of the log NC)
$U(z_i)$	natural statistic, first component in pairwise interaction model
V_+	In the Ising model, the number of positive spins
V_{++}	In the Ising model, the number of neighbouring positive spins
$V(z_i, z_j)$	natural statistic, second component in pairwise interaction model
$V(z)$	natural statistic for distribution in exponential family
X	covariates impacting on data Y
Y	observed data
Z	Underlying spatio-temporal process
z_i	observed value of Z at spatio-temporal location i
$i : +\delta$	Denotes the δth nearest neighbour. *E.g.* The nearest neighbhours of site i are denoted by $i : +\delta$ with $\delta \in \{+1, -1, +2, -2\}$ respectively referring to neighbours to the east and west, and south and north. See page 67 for more details.
$\mathcal{B}$	the set of all possible events or subsets of the sample space Ω
$\mathcal{C}$	the collection of all cliques on the sampling window as defined in Section 4.2.2.
$\mathcal{H}$	The Hamiltonian or Energy in a Gibbs random field, Section 4.2.4.
$\mathcal{L}$	lattice index set denoting range of subscripts i of z_i; $\mathcal{L} = \{i : i = 1, 2, \ldots, L\}$
$\mathcal{N}$	neighbourhood basis as defined on page 67
$\mathcal{N}_k$	subset of elements in a neighbourhood basis as defined in Section 4.2.2.
$\mathcal{N}$-nbrhd	neighbourhood defined by $\mathcal{N}$.
$\Re^+$	set of positive real numbers
$\mathcal{T}$	Temperature in a Gibbs random field of Section 4.2.4
$\mathcal{X}$	parameter space of parameter x in discussion of MCMC simulation approach in Section 4.4
$< \Omega, \mathcal{B}, \mu >$	A random field incorporates the sample space, all possible events, and the probability measure.
α	effects of covariates
δ	label of neighbour, denoting its position with respect to the central site, as defined in Figure 4.2
Δ	denotes the site of neighbouring site labels, $\delta \in \Delta$, where *e.g.* $\Delta = \{-1, +1, -2, +2\}$ for a first order neighbourhood.

continued on next page

Table 1: (continued from previous page)

Notation	Description
$\gamma_+^{\mathcal{C}}$	Used in Braggs-William approximation to the NC, reflects global marginal proportion of presence, and represents long-range dependence
$\gamma_{++}^{\mathcal{N}}$	Used in Braggs-William approximation to the NC, reflects local conditional proportion of presence, and represents short-range dependence
γ	Used in Braggs-William approximation to the NC, value of $\gamma_{++}^{\mathcal{N}}$ that maximises a contribution to the NC.
κ	indices of covariates
$\lambda(A, B)$	ratio of normalization constants based on parameters θ_A and θ_B
μ	ordinarily the mean of Y given Z
μ	also the probability measure defined on $\mathcal{B}$ in Section 4.2.4.
$\pi(x)$	equilibrium distribution (distribution of interest) in discussion of MCMC simulation approach in Section 4.4
$\phi(\theta)$	component of $h(z, \theta)$ involving only θ
ϕ	In statistical physics, this represents the fugacity in the Ising model
$\psi(z)$	component of $h(z, \theta)$ involving only z
τ	index to components of q
η	link function linking mean μ, covariates X and effects α
θ	spatio-temporal dependence parameter in distribution of Z
Ω	Configuration space for spatio-temporal process Z
Ω_0	$\{0, 1\}$
AL	Autologistic distribution (Chapter 4)
CAR	Conditional autoregressive spatial model (Cressie 1993)
EM	Expectation-Maximization computational algorithm (Dempster, Laird & Rubin 1977)
GAM	Generalized additive model
GLM	Generalized linear model (McCullagh & Nelder 1993)
IMCS	Integrated mean canonical statistic
ISMC	Importance sampling Monte Carlo
MC	Monte Carlo
MCMC	Markov chain Monte Carlo
MLE	Maximum likelihood estimation
MRF	Markov Random field
NC	Normalization constant
pdf	probability distribution function

Statement of Original Authorship

The work contained in this Thesis has not been previously submitted for a degree or diploma at any other higher education institution. To the best of my knowledge and belief, this Thesis contains no material previously published or written by another person except where due reference is made.

Signed: QUT Verified Signature

Date: 6 April 2001

Acknowledgements

Grateful thanks to the following people
for their support during my thesis.

Supervisor, Professor Tony Pettitt
Enthusiastic
guide with infinite patience
down a complex path.
Subtle teacher in
practice and philosophy
applying statistics.

Shane.
Near my side through this
adventure as scribe, chauffeur,
Coco and partner.

Low Choy and Clouston families, and many friends:
Keeping me grounded;
lending perspective and faith,
and washing machine.

Fellow postgrads: esp. Shelley, Peter, Bear, Angie, Jodie, Neil, Fiona.
Friends who shared with me
these roller coasters of life
and of intellect.

IT support: Neil and Mick.
Calm stable platform,
streamlined bibliography,
day and night access.

Teachers at the University of Queensland:
Drs Chant, Pollett, Smythe, MacGillivray,
and the late Professor Steve Lipton.
Enticed me into
statistics, revealing a
beauty and richness.

The Australian Research Council.
Financial support:
Australian Postgraduate
Research Award.

The School of Mathematical Sciences,
Queensland University of Technology.
Financial support:
University visits
and Conferences.

University of Bristol, esp. Professor Peter Green.
Hospitality,
Markov chain Monte Carlo,
and some seed ideas.

Colleagues at the Environmental Protection Agency,
and Queensland Parks & Wildlife Service.
For sharing milestones,
gentle encouragement and
flexible schedule.

Dedication

To my parents, for nurturing me with tangrams, peg-counting, encyclopaedias and dolls with randomized conversations.

To my mother, Kim Low Choy, for inspiring me with her pioneering spirit and her appreciation of knowledge and beauty.

To my late father, Warren Radford Low Choy, for inspiring me with his creativity and love of nature.

To my twin brother and sister, Daniel and Jodie Low Choy, for their optimism.

And most of all to my partner and fiancé, Shane Clouston, who has never known me *sans* PhD, and has been a bedrock of good humour, love, and a unique mixture of support and challenge.

Chapter 1

Preamble

Contents

1.1 Introduction 2
1.2 Scope of thesis 2
1.3 Outline of thesis 5
1.4 Contributions 6

1.1 Introduction

> It began in mystery, and it will end in mystery, but what a savage and beautiful country lies in between.
> *-Diane Ackerman (b. 1948)*

In this thesis I investigate a problem in applied statistics. Theory is developed specifically to address the needs of a particular combination of data and research questions, which arise quite naturally in biogeographical applications as well as the analysis of computer and satellite imagery, including remote sensing data. Throughout the thesis I use a particular case study on dingo behaviour to motivate and illustrate theory.

I focus on the analysis of binary data observed on a two-dimensional lattice. Binary data may be coded as zeroes and ones representing any pair of opposites, for example: absence and presence, failure and success, or magnetism spins up and down. A difficulty that can occur with this type of data is that recording a zero can be ambiguous: it can represent either a true zero or a missing value. For example, a map of success and failure for finding a bird species may contain zeroes which represent true success (no birds found at that location given some research effort) or, alternatively, zeroes may represent missing values (no research effort at that location). In this case there is an underlying process for research effort which is confounding the observations of success and failure for locating the bird species.

One way to clarify this ambiguity is to follow a common approach in situations like this and construct a hierarchical model, retaining the model for the observations of success and failure, but also introducing an underlying presence/absence process into the modelling framework. Thus zero responses can arise either from a failure (no bird species located) in the original process given presence (some research effort at the location) modelled by the underlying presence/absence process, or alternatively a zero response can arise from a necessary failure (could not locate any bird species) due to absence (no research effort at the location) modelled by the underlying presence/absence process.

This approach raises many modelling and inference issues: relevant and motivating applications; modelling binary data observed on a two-dimensional lattice; modelling underlying spatio-temporal dependence; hierarchical modelling; frequentist or Bayesian approach to inference; computational issues encountered with a Bayesian approach. The next section (1.2) outlines the scope of this thesis by expanding on each of these issues.

Once the scope has been established, this introduction finishes by outlining the structure of the rest of the thesis (Section 1.3); and by highlighting contributions made by the thesis (Section 1.4).

1.2 Scope of thesis

Case study on dingo behaviour

This research was indirectly started when an agricultural scientist (Mitchell 1988) asked a specific question on **dingo behaviour**: "Which of these chemicals are dingoes most attracted to in the wild?" The corollary was "Of these highly attractive chemicals, are any successful enough in practice to warrant widespread use for management of dingo populations?". To answer these questions, a field experiment was designed and data collected according to accepted statistical methodology (Pettitt & Low Choy 1999). Chemicals were arranged in pairs on a two-dimensional space-time grid. Dingo visits to each chemical were

recorded as a binary response. Inspection of the data revealed that there were large patches in time and along the transect when no dingoes were observed to be attracted to any chemicals. Thus it was difficult to distinguish between zero responses which arose due to an absence of dingoes, compared to a lack of response although dingoes were present. This prompted an extension of the model to incorporate an underlying presence/absence process.

Presence/absence processes are common in the areas of biogeography, where they are used to map presence of various species of fauna and flora, and in the area of image analysis, where they can be used to model black and white pixellated images obtained from cameras, satellites, radar or medical imaging devices. These applications also suffer from the problem of ambiguous zeroes, which may represent either missing data or zero response. Thus examination of this case study lends itself to wider applications.

Underlying spatio-temporal dependence

Statistical methodology is well-developed in the area of independent data, indeed most newcomers to statistics are often only exposed to methods suitable for independent data (Walpole, Myers & Myers 1998, Siegel & Castellan, Jr 1988). Methodology for time-dependent data is also well-developed, beginning early in the twentieth century with auto-regressive and moving average models. There is a wide variety of approaches available today. A few of the most common are auto-regressive integrated moving average (ARIMA) models (Box & Jenkins 1970) and extensions, spectral time series approaches (Priestley 1981), Dynamic Linear models (West & Harrison 1997). The next logical step has been to expand to the arena of spatial or **spatio-temporal dependence** between observations. This thesis forms a small part of a growing interest (Upton & Fingleton 1990, Diggle 1983, Ripley 1988, Cressie 1993, Diggle, Tawn & Moyeed 1998) in statistical modelling for **spatio-temporal dependent** data which has recently made rapid progress due to increased computational power.

In this thesis, I concentrate on situations where the spatio-temporal dependence is not central to the research question of interest but rather is **underlying** to the problem. This occurs, for example, when estimating treatment effects from a designed experiment with underlying spatial dependence between observations (Chakraborty, Pettitt, Boland, Low Choy, Cameron, Irwin & Davis 1993, Chakraborty, Pettitt, Low Choy & Boland 1995, Pettitt & Low Choy 1999); or when predicting the true image given an observed image where neighbouring pixels are known to be similar (Geman & Geman 1984, Besag 1986, Dubes & Jain 1989, Weir & Pettitt 1999).

Nonetheless, the techniques and models developed in this thesis may be applied to situations where the spatio-temporal dependence *is* central to the research question. This may occur, for example, in pattern analysis or texture recognition where the researcher wishes to summarise a spatial pattern or texture by a small set of parameters. These parameters can be used as indicators for comparing two populations of presence/absence data to determine whether they are similar.

Binary two-dimensional lattice data

Spatial datasets fall into several main categories which depend on the way in which the spatial attributes of each measurement are recorded: for example, as grids or as points, polygons/regions, networks. In the image analysis literature gridded data is known as raster data and the others are various forms of vector data. It is often advantageous to transfer

between raster and vector formats depending on the spatial operation being applied to the data. There is a wide research area as to how best to achieve this transformation (Dubois 2001). For example, polygon data may be transformed to gridded data by aggregation or dis-aggregation. The focus of this thesis is on gridded data and thus, by applying the logic above, by extension to other types of spatial data. In the statistical literature, gridded data is more generally known as lattice data. A particular focus of this thesis is **two dimensional lattice data**.

The most common methods used to analyze spatial dependence of two dimensional lattice data include Markov random field (MRF) models such as auto-Gaussian and auto-Poisson models, (Besag 1986, Cressie 1993, Wolpert & Ickstadt 1995) and spatial linear models based on Gaussian error structures. These models are not suitable for binary response data. In this thesis we consider the case where there is a **binary** response, for example presence/absence and success/failure data (Cox 1970).

Hierarchical model

The spatial data types mentioned above correspond directly to coverage types in geographical information systems (GISs) such as ARC/INFO (Environmental Systems Research Institute 1997), ArcView (Environmental Systems Research Institute 1996), and MAPINFO (MAP INFO Corporation 1997); or image analysis packages such as Imagine (ERDAS 1998). Currently GISs are strong in spatial data collation, management, visualization and complex spatial queries. They have the capacity for "spatial analysis" which in this context mostly corresponds to mathematical and topological operations on spatial datasets, which may for example use information on aspect, elevation and slope to enhance spatial data. There is little capacity in these systems, however, for statistical modelling apart from limited kriging. The image analysis packages on the other hand permit many different methods of image filtering to remove white noise from a "dirty" image, in the tradition of Besag (1974).

There are many geostatistical packages (Dubois 2001) available, to address geostatistical requirements of interpolating a surface given point observations. These packages offer a variety of kriging methods, which rely on modelling a spatial variogram. Other statistical packages, of which Splus (Becker, Chambers & Wilks 1988) and SpaceStats (Anselin & Smirnov 1998) are the major ones, address spatial statistical modelling for lattice data. These models include spatial exploratory data analysis, spatial extensions to ARIMA and spatial general linear models. This thesis investigates a **hierarchical model** not available in these packages suitable for incorporating treatment effects of covariates and **underlying presence/absence** for **binary data on a 2D lattice**.

By using a **hierarchical model** inference follows intuition and can be tailored to the hierarchical structure of the data and latent variables. In this thesis hierarchical models fit well with the Bayesian approach (Lindley & Smith 1972), and assist in organizing a complex model with many interactions between entities and various sources of error.

Within a layer of the **hierarchical model**, a binary Markov random field is used to model the **underlying spatio-temporal presence/absence process**. This provides a wide choice of models, including the Ising model (Ising 1925) and the closely related Autologistic model (Besag 1974), Verhagen's model (Verhagen 1977) and Pott's model (*e.g.* Georgi (1996)). I pay particular attention to the three-parameter autologistic model, which is novel in spatial statistics contexts where the single-parameter autologistic model appears more popular. Additional parameters included account for overall prevalence and change the assumption of isotropy of spatial dependence to allow for different amounts of spatial

dependence in two directions.

Bayesian and Frequentist approaches

Implementation of the modelling led me to explore both **frequentist** and **Bayesian** approaches. Exploration of application of both statistical philosophies to the same problem served to highlight the strengths of each approach.

A refinement to current frequentist approaches to analysis of such data resulted in application of the EM algorithm to inference and incorporation of spatial information via blocking or smoothing. To fully implement the **Bayesian** approach to modelling and inference I embarked on a theoretical examination of new computational techniques to estimate a normalization constant.

A major challenge of the **hierarchical** approach in the **Bayesian** context is the estimation of a normalization constant for the underlying binary Markov random field. This has been a long-standing problem in the area of statistical physics (referred to as the "Ising" disease by one researcher) and has not been solved for all binary Markov random fields, including the three-parameter autologistic, although it has been solved for the single-parameter Ising version (Domb & Green 1972*a*).

1.3 Outline of thesis

Chapter 2 presents two literature reviews. The first (Section 2.3) motivates this research by identifying suitable applications, including the *dingo* case study. This case study is based on a field experiment investigating the attractiveness of chemicals to wild dingoes (Section 2.3.1). These motivating examples and case study provide an important vehicle for illustrating concepts throughout the thesis, and are all potential candidates for applications of the method. Other applications include prediction of the true image given ground truthed observations and radar imagery of presence/absence of various flora species in forestry applications, for instance prediction of presence/absence of white cypress pine (Section 2.3.2). Another application is to estimation of presence/absence of toads in Finland given the underlying degree of research effort (Section 2.3.3). The increasing availability of huge repositories of digital imagery, such as that provided by satellite photography or images, provides a large pool of potential uses for the methods investigated in this thesis.

The second literature review in Chapter 2 summarizes models potentially suitable for modelling binary two-dimensional lattice data with ambiguous zeroes and underlying spatio-temporal dependence.

The major part of the thesis is devoted to investigating alternative approaches to inference for binary data observed on a two-dimensional lattice. Both frequentist and Bayesian approaches to inference are investigated.

The first approach investigated is frequentist (Chapter 3). This chapter is comprised mostly of a published paper, written with the supervisor (Pettitt & Low Choy 1999). The paper uses a frequentist approach employing the EM algorithm together with bootstrapping to estimate standard errors of parameter estimates. A novel hierarchical model was constructed and inference tailored to the case study by using the EM algorithm. Spatial (one-dimensional) and temporal variation in underlying presence at each experimental unit, or block of units, was modelled using blocking and smoothing techniques. These were based on an assumption of independence between experimental units. This did not fully capture the underlying spatial and temporal dependence of the presence and absence of dingoes, and

led to examination of Bayesian alternatives. In addition the Bayesian treatment was seen to have other advantages, primarily in interpretation, and in provision of the full posterior distribution of parameters compared to interval estimates based on bootstrapped standard errors.

Chapter 4 provides a literature review of the background required for the theoretical aspects of a Bayesian approach to modelling. The family of binary MRF models, especially the three-parameter autologistic model, is defined and its properties examined. *The in-depth treatment afforded to the three-parameter autologistic model is, I believe, new work.* Since analytic examination of the models is difficult, inference for these models hinges on the ability to simulate from them. This opens discussion on MCMC methods, which were invented to simulate from a binary MRF, the Ising model. At the time that the thesis was started there was little published on these methods, however there is now a wide body of literature in the area. MCMC methods, their principles and associated diagnostics, are therefore outlined briefly. This chapter continues with an overview of existing inference methods suitable for estimating (distributions of) parameters in binary MRFs. Existing methods all make assumptions not suitable for the case study which is the focus of the thesis.

A Bayesian approach was undertaken in gradual stages. A preliminary investigation of a three-tier hierarchical model using a Bayesian approach is given in Chapter 5. *This form of hierarchical model is new and a contribution of the thesis.* The application of the model to this type of experiment is also novel. Before embarking on a full Bayesian analysis based on an extended four-tier hierarchical model in Chapter 7, a theoretical problem had to be addressed.

The theoretical problem separating the three-tier and extended four-tier models is how to estimate a normalization constant for the three-parameter autologistic model. This was recognized to be a special case of a more general problem studied in Chapter 6: that of estimating a normalization constant for exponential family distributions for binary data. One approach in particular was found to be successful, and can be described as an estimating equations approach, the integrated mean canonical statistic and is known as the path sampling method (Gelman & Meng 1998). *Technical issues of its implementation for the three-parameter autologistic model have been addressed for the first time, I believe, in this chapter.*

In Chapter 7 I fully address the applied statistical problem of using inference for binary Markov random field models for the *dingo* case study. The extended four-tier model is used. *This is a new fully Bayesian approach to inference with a hierarchical model for this type of problem.*

Discussion and conclusions are given in the final Chapter 8. Further extensions of the work begun in this thesis are also indicated in the final chapter.

1.4 Contributions

Contributions made by the thesis to the field of applied statistical modelling can be listed under general headings for the application and for statistical methodology as follows.

Environmental management and behaviour of dingoes

The thesis contributes to the field of application, and by extension to other fields of application, by demonstrating the application of the methods to real data. Contributions to the

application area include:

- serious treatment of the shortcomings of available models for analysing the *dingo* case study data;
- case study of application of hierarchical modelling to presence/absence data with underlying spatial (temporal) dependence;
- comparison of frequentist and Bayesian approaches to solution of real problem
- a complete answer to the research question "How well do these chemicals attract dingoes in the wild?"

Spatial statistical modelling for binary data on 2D lattice

The thesis also contributes to the methodological issues of inference for binary Markov random fields as well as spatial statistical modelling, specifically for binary data on two-dimensional lattices. Contributions to statistical methodology include:

- identification of a specific modelling problem that arises in areas such as field experiments with underlying spatial variation and analysis of computer generated imagery;
- development and demonstration of a viable alternative to the single parameter Ising model for spatial statistical modelling—the two- or three-parameter autologistic models;
- integration and synthesis of information on binary Markov random field models from the diverse subject areas of statistical physics, image analysis and spatial statistics;
- development of an EM algorithm to obtain point estimates and bootstrapping to obtain standard errors for parameters in the hierarchical model with simplified spatial modelling;
- development of methods for, examination and demonstration of practical issues involved in, estimation of normalization constant for members of the exponential family using the path sampling method;
- development of a MCMC algorithm to obtain posterior distributions of parameters in a three-tier hierarchical model with a three-parameter autologistic model capturing underlying spatial (temporal) dependence;
- investigation of two extensions to this hierarchical model in a Bayesian context;
- implementation of MCMC diagnostic methods suitable for discrete data.

Two publications have resulted from work contained within this thesis. Pettitt & Low Choy (1999) covers Chapter 3 by providing the frequentist approach to analysis of a restricted hierarchical model. The second publication (Low Choy & Pettitt 1997) presented the work from Chapter 7. This was presented as a poster at the Bayesian Statistics 6 conference of the Royal Statistical Society in July 1997, and was awarded first prize amongst all the submitted posters (numbering over 100) by the judge Alan Gelfand.

Chapter 2

Motivation

Contents

2.1 Introduction 10
2.2 Zero-inflated data 10
2.3 Motivating applications 12
2.3.1 *Dingo* case study 12
Design 13
Statistical challenges 16
2.3.2 Applications to Flora distribution maps: *Cypress* application . . . 17
2.3.3 Biogeographical applications: *Toad* case study 20
2.4 Overview of models for binary spatial data 23
2.4.1 Multivariate techniques 23
2.4.2 Aggregation 23
2.4.3 Generalized Linear models 24
2.4.4 Markov Random field models 25
2.4.5 Extrapolation of time series models 26
2.4.6 Stochastic processes 27
2.4.7 Geostatistical models 28
2.5 Discussion 28

2.1 Introduction

Presents, I often say, endear Absents.
-Charles Lamb Essays of Elia (1823) 'A Dissertation upon Roast Pig'

The research in this dissertation was motivated by a problem encountered when applying statistics to the analysis of a real scientific question where there are several methodological issues which require further analysis. This chapter details these motivating factors, especially the problem of zero-inflated data (Section 2.2) and the *dingo* case study (Section 2.3.1). It also describes other applications where this problem might arise (Section 2.3) and provides a brief literature review to provide some background in methods and models that might be considered appropriate for binary spatially dependent data (Section 2.4).

2.2 Zero-inflated data

With spatial presence/absence data, a zero can represent either a zero response or a structural zero resulting from missing data. The problem of ambiguous zeroes which is commonly encountered with presence/absence data also occurs in the context of count data, where it is referred to as the problem of inflated or extra zeroes. Models are available for handling count data where an underlying process is producing more zeroes than expected under a particular modelling framework. A threshold effect can govern an underlying presence/absence process, where presence corresponds to triggering the threshold. Once presence occurs counts can then be observed. Thus a zero response can result from absence or else a zero count given presence.

Researchers have chosen different approaches to dealing with these ambiguous zeroes for count data. Early work by Aitchison (1955) used a parametric mixture model with a point mass at zero.

Lambert (1992) investigates zero-inflated Poisson (ZIP) regression applied to the analysis of incidence of defects in a manufacturing process. In this situation it was hypothesised that "slight, unobserved changes in the environment cause the process to move randomly back and forth between a perfect state in which defects are extremely rare and an imperfect state in which defects are possible but not inevitable." This perfect state therefore corresponds to being below the threshold described above. See references contained in this article for other applications and related methods for dealing with zero-inflated data.

This author extends ZIP models without covariates (Johnson & Kotz 1969) so that both the mean rate λ and the probability p of being below the threshold can depend on covariates. Assuming responses $y = (y_1, \ldots, y_n)^\top$ are independent, the ZIP model is

$$y_i \sim \begin{cases} 0 & \text{with probability } p_i \\ \text{Po}(\lambda_i) & \text{with probability } 1 - p_i \end{cases} \tag{2.1}$$

and parameters are modelled as

$$\begin{aligned} \log(\lambda) &= B\beta \\ \text{logit}(p) &= G\gamma \end{aligned} \tag{2.2}$$

where B and G are matrices of covariates for the mean rate and threshold probability respectively; β and γ are vectors of corresponding parameters; and $\log(\lambda)$ is a vector with

elements $\log \lambda_1, \ldots, \log \lambda_n$, similarly for logit (p). The probability distribution can therefore be written:

$$\begin{aligned} p(y_i = 0) &= p_i + (1 - p_i)e^{-\lambda_i} \\ p(y_i = k) &= (1 - p_i)e^{-\lambda_i}\lambda_i^k \Big/ k!\,, \quad k = 1, 2, \ldots \end{aligned}$$

Maximum likelihood estimates of p, λ, β and γ were estimated via the EM algorithm (Dempster et al. 1977). Iteratively reweighted least squares (Green 1984) was used to provide iterative solutions.

A more general option is to use mixtures of discrete distributions (Böhning 1999) such as zero-truncated Poisson distributions,

$$q(r, \pi, \lambda) = \sum_{j=1}^{k} \pi_j P_0(y, \lambda_j(x))$$

where $\lambda_j(x)$ can be further parameterized as $\exp\{\alpha_j + \beta^{\mathsf{T}} x\}$.

The hurdle model of Mullahy (1986) or conditional model of Welsh, Cunningham, Donnelly & Lindenmayer (1996*a*) is a hierarchical model for responses y given a covariate z. The general form is (Welsh et al. 1996*a*):

$$\begin{aligned} p(y = 0|z) &= 1 - p(z) \\ p(y = k|x, z) &= p(z)q(k, \theta(x)), \quad k = 1, 2, \ldots \end{aligned} \tag{2.3}$$

where q is a probability distribution on $\Re^+$ such as the truncated negative binomial, and $\theta(x)$ is a vector of parameters which may depend on x. Here the distribution q can be interpreted either as the conditional distribution of y given $y > 0$ or else as a mixture of two distributions defined over disjoint supports. The parameters p and θ are separable in this formulation of the model. Heuristically, fitting can proceed by first estimating $p(z)$ via use of a Generalized Linear model (GLM) (McCullagh & Nelder 1993) with link function suitable for binary data and response $I[y_i > 0]$. Secondly, $q(\cdot)$ is estimated conditional on $y_i > 0$.

The model can be extended (Welsh et al. 1996*a*) so that the link between covariates and p and λ are in the form of the link functions of Generalized Linear models (GLMs) or Generalized Additive models (GAMs), *e.g.*

$$h(p(z)) = z^{\mathsf{T}}\beta \quad \text{or} \quad h(p(z)) = \sum_{j=1}^{r} l_j(z_j)$$

An option with GLMs is to specify overdispersion of the variance within a Binomial or Poisson distribution (McCullagh & Nelder 1993). For a response variable y, if the mean function is $\mu = \mathrm{E}[y]$ then overdispersion can be specified by choosing α such that:

$$\mathrm{Var}[y] \propto \mu^{\alpha} \quad 1 < \alpha < 2$$

A quasi-likelihood approach, in contrast to GLMs, specifies only the mean and variance functions instead of the whole distribution (McCullagh & Nelder 1989). Chiou & Müller (1998) consider a regression model with mean $\mathrm{E}[y] = \mu$, link g to explanatory variables x given by $\mu = g(x^{\mathsf{T}}\beta)$, and variance function $\mathrm{Var}[y] = \sigma^2(\mu)$. They allow the link and variance functions to be unknown but smooth using semi-parametric models.

Aitchison & Ho (1989) construct a multivariate Poisson-Log Normal distribution where a response y is modelled as a mixture of d independent Poisson variates with rates λ_j. The latent rates λ can arise from a multivariate log Normal distribution with mean μ and variance-covariance matrix Σ. Due to the independence of the responses from each group, the likelihood $p(y|\ldots)$ may be factorized as $\prod_i p(y_i|\ldots)$ and the probability density integrates out the latent parameters λ

$$p(y|\mu,\Sigma) = \int_{\Re_+^d} \prod_{i=1}^{d} p(y_i|\lambda_i) g(\lambda|\mu,\Sigma) d\lambda \quad y_i = 0, 1, \ldots \tag{2.4}$$

Estimation proceeds first by reparameterization of Σ to a lower triangular form T and then applying maximum likelihood using Newton-Raphson and steepest ascent methods. Transforming the integrals to Hermitian form improves computations. This gives a hierarchical model with latent parameter vector λ.

Thus there are a variety of ways to handle extra zeroes. The modelling approach of Lambert (1992), and some aspects of Welsh et al. (1996*a*) and Chiou & Müller (1998), will be applied to analysis of the *dingo* case study. When the observed data are also binary, *e.g.* success/failure data, then a Bernoulli model can be used for the data layer. The underlying spatio-temporal dependence may be modelled by allowing the probability of presence to vary over blocks or smoothly over sites (Chapter 3). Alternatively an underlying spatio-temporal dependence process may be modelled by a binary Markov Random Field (MRF) such as the autologistic model. A hierarchical Bernoulli-autologistic model is analyzed in both frequentist (Chapter 3) and Bayesian contexts (Chapters 5–7) in this thesis.

2.3 Motivating applications

There are many potential applications for the work in this thesis. The focal case study used to illustrate theory throughout this work is the *dingo* case study presented below in Section 2.3.1. It is an example of a wider problem in the analysis of field experiments with underlying spatial variation and so-called zero-inflated data, briefly reviewed in Section 2.2.

Other major applications arise in the area of biogeography, where distribution maps (population atlases) of flora and of fauna are of interest. Applications in forestry (Section 2.3.2) have previously been addressed as a highly multivariate problem, with numerous covariates and therefore problems with collinearity. Estimation of a population atlas of toads in Finland (Section 2.3.3) raises generic questions for population atlases of any species. These maps depict estimated presence and absence of species over gridded spatial areas. Absence in these cases can represent absence or missing data. These are multidisciplinary, as both involve interaction between biologists and computer scientists producing computer generated imagery.

However this work should find application to a variety of application areas such as those described later in the literature review of models for two dimensional binary lattice data (Section 2.4).

2.3.1 *Dingo* case study

Many plot and field experiments predominate in, but are not restricted to, the agricultural, environmental and biological sciences. An omnipresent feature of data collected in this way is spatial dependence between observations.

As sheep meat and wool contribute significantly to Australia's economy, much effort has therefore been expended in environmental management and control of predators of sheep such as the dingo and feral dogs. Various methods of control for these animals rely on luring and entrapment, so success of lures is of paramount importance and scientific interest. A series of experiments described in Mitchell (1988) was prompted by a lack of scientific knowledge of dingo behaviour which might affect the success of these lures, in particular the dingoes' sense of smell. This lack of knowledge extended to dingo communication relating to territory, social status and mating, which also revolves around their olfactory sense. In these pilot experiments comparison was made of several recipes for chemical lures obtained from experience with coyote management or from local folklore and tradition.

Pen trials refined a number of chemical lures which were believed to have potential as lures. A series of field experiments was used to validate the results of the pen trials. This round of field experiments was restricted to eight chemicals and a control, and again covered areas of low, medium and high dingo density, at various seasons covering important stages of the dingo life cycles: whelping (August), denning (November), dispersal of juveniles (February) and mating (May).

Dingo exposure to lures and their visitation rates were affected both by seasons and by population density which affected their propensity to search widely. In addition attractiveness of different lures was affected by both the seasons and population density, which changed predominant behaviour patterns.

The *dingo* case study presented here was also based on a field experiment, focussing on just six chemical lures, in a relatively high population density area during just one season.

The objectives of the series of dingo experiments culminating in the case study of interest therefore focussed on environmental control issues. Most important was the identification of the most effective chemical attractants. Then, after the logistics of using lures to attract dingoes to particular sites was optimized, specificity of lures was needed to ensure deterrence of non-target species, such as domestic animals and protected species. Finally the choice of chemical attractants was to be incorporated into guidelines for efficient and dingo-specific environmental management practices.

Design

Transect sampling is common in animal and plant abundance studies (Cressie 1993, for example) since it provides a practical and resource efficient way of sampling a large sample area, whilst collecting information on spatial dependence at a range of spatial scales. Thus the design of the experiment was structured to lie along a transect. A remote 100 km section of a track provided a convenient ready-made transect. Daily variations were dealt with by obtaining observations every day for a week. This provided a two-dimensional grid of observations, one dimension being spatial position along the transect, and the other being temporal (day within week).

An important aspect of dingo behaviour to be considered was the proportion of a given area covered by dingos in a given day. Corbett (1995) suggests that dingos tend to travel widely from day to day in small family packs, covering 10–15% of an area when there is a *good* concentration of dingos. Thus a fairly low prevalence of dingos was expected (with 10–15% being an optimistic estimate). Placing more than one chemical at a site would therefore maximise the power of the information collected. In this experiment, *different* chemical types were placed at two *locations within a site.* This facilitated pairwise comparisons between the attractiveness of chemical types.

In order to avoid any confounding between the order of the chemicals placed along this transect, the chemicals were grouped together in blocks consisting of one each of the different types of chemicals. Spatial variation could thus be partially accounted for. The final design concentrated on incorporating some balance between neighbouring chemicals, thus ensuring that chemicals would not be situated near any other chemical more often than others.

The design that was used was based on blocks, each consisting of 3 sites situated 500m apart, and 2 locations within a site situated 50m apart, as depicted in Figure 2.1. The distance between blocks was the same as the distance between sites (500m), so blocks were merely a means of ensuring that every chemical type was represented in every 1650m section of the transect. Each site therefore constitutes a 'plot' within a block of 3 sites and a location is a 'subplot' within a 'plot'. The 'plot' is balanced since it contains all treatments, whereas the 'subplot' contains only two such treatments, and is therefore an incomplete block.

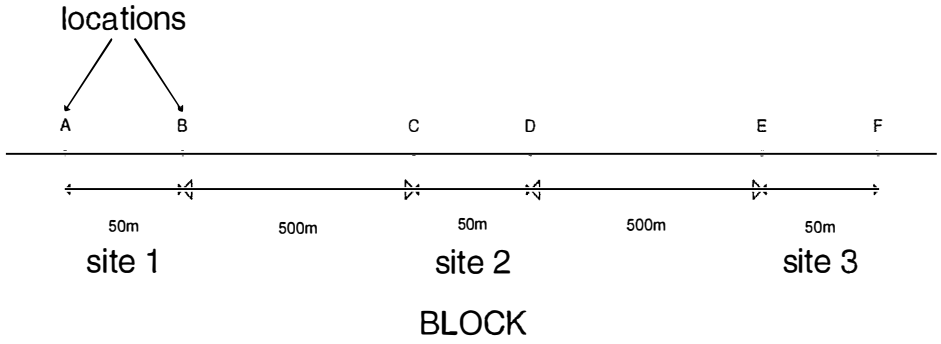


Figure 2.1: Design: arrangement within a block

There are therefore $1 \times 5 \times 1 \times 3 = 15$ different combinations of chemicals within a block, disregarding order of sites within block, and locations within sites. Eg

$$\begin{array}{lll} AB & CD & EF, \\ \hline AB & CE & DF, \\ \hline AB & CF & DE, \\ \hline AC & BD & EF, \\ \hline \ldots & \ldots & \ldots \end{array}$$

Arrangement of chemicals at locations within sites, positioning of sites within blocks, positioning of blocks within replications were randomised to eliminate any systematic variation due to these orderings.

Three replications of all fifteen arrangements of blocks were considered feasible. This gave $3 \times 15 = 45$ blocks, $45 \times 3 = 135$ sites, and $135 \times 2 = 270$ locations. With this design, each particular pair of chemicals appears only 9 times.

The absence or presence of a visit by a dingo at each location within a site at each location was recorded for each day, for the length of the transect. The presence of a dingo in the near vicinity of the chemical was indicated by footprints or urine or other evidence

of a dingo being close to that location. Consequently, it was not possible to determine how many dingos visited the site on a given day, or how many times a particular dingo visited the site on a given day. Observations were binary:

$$y_{vst} = \begin{cases} 1, & \text{if any dingo visited location } v \text{ at site } s \text{ on day } t \\ 0, & \text{otherwise} \end{cases} \tag{2.5}$$

The same arrangement of chemicals was used each day. Once each day, to minimize carry-over effect, all evidence of dingo visits were removed and chemicals replaced.

This experimental design was repeated for eight days. However, the data collected on the first day was discarded as it was incomplete (only two-thirds was collected), and used as a training exercise in data collection methodology. The complete design is shown in Table 2.1.

Table 2.1: Design: dingo experiment. Tabulation of chemical types.

Replication 1		
	location	
site	1	2
1	C	D
2	A	F
3	B	E
4	B	C
5	F	A
6	D	E
7	A	C
8	B	E
9	D	F
10	A	D
11	B	C
12	F	E
13	D	B
14	A	E
15	F	C
16	A	D
17	C	F
18	B	E
19	E	F
20	C	A
21	D	B
22	D	C
23	E	A
24	F	B
25	C	F
26	B	A
27	E	D
28	A	F

Replication 2		
	location	
site	1	2
46	A	C
47	B	E
48	F	D
49	F	E
50	C	B
51	A	D
52	E	A
53	F	D
54	B	C
55	F	B
56	C	E
57	A	D
58	D	C
59	F	A
60	B	E
61	B	F
62	C	D
63	A	E
64	F	C
65	E	A
66	B	D
67	D	E
68	B	A
69	C	F
70	E	F
71	C	A
72	B	D
73	C	D

Replication 3		
	location	
site	1	2
91	F	A
92	C	B
93	E	D
94	D	A
95	B	F
96	E	C
97	E	C
98	D	F
99	B	A
100	D	C
101	B	A
102	F	E
103	E	F
104	D	A
105	C	B
106	C	D
107	E	A
108	B	F
109	B	C
110	D	F
111	A	E
112	C	F
113	B	D
114	A	E
115	F	A
116	B	E
117	D	C
118	B	E

continued on next page

Table 2.1: (continued from previous page)

Replication 1		
	location	
site	1	2
29	E	C
30	B	D
31	F	D
32	E	A
33	B	C
34	B	A
35	E	C
36	D	F
37	B	A
38	D	C
39	E	F
40	A	C
41	D	E
42	F	B
43	F	B
44	D	A
45	C	E

Replication 2		
	location	
site	1	2
74	E	F
75	A	B
76	D	E
77	A	C
78	B	F
79	A	D
80	B	E
81	F	C
82	B	D
83	E	C
84	A	F
85	C	E
86	F	D
87	B	A
88	E	D
89	F	A
90	B	C

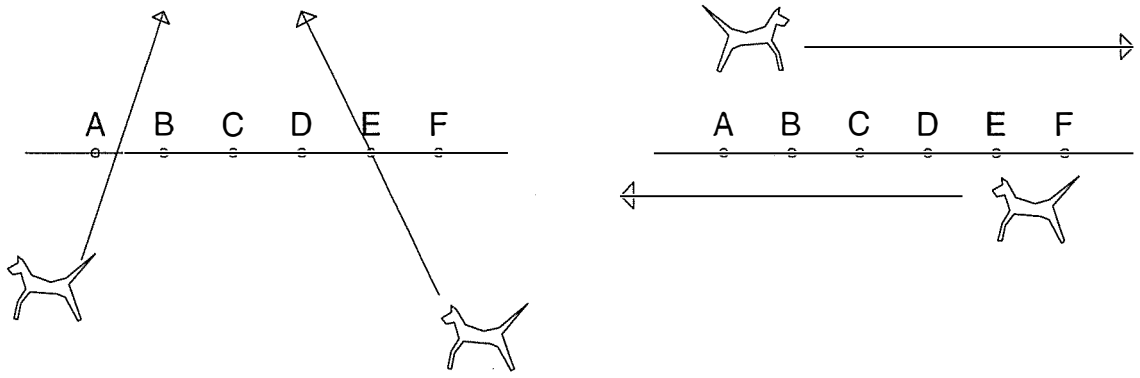
Replication 3		
	location	
site	1	2
119	C	F
120	D	A
121	D	B
122	E	F
123	A	C
124	A	C
125	E	B
126	F	D
127	C	F
128	E	D
129	B	A
130	E	D
131	A	C
132	B	F
133	E	C
134	F	A
135	B	D

Since this experiment was quite new, scientific expert knowledge on dingo behaviour was canvassed to inject vital information into the modelling process. There were some issues raised whilst applying this knowledge. The two biologists requiring the analysis each had completely different ideas about the manner in which dingos covered an area such as a road. The first biologist believed that the dingos would be more likely to cross a road as though it were not there. This would mean there would be little, if any, spatial dependence between observations collected along a transect based on a road. The second biologist expected the dingos to follow a road, since it would provide a convenient medium for traversing their territory quickly. In this case, a relatively large amount of spatial dependence between observations taken along a road-transect would be expected. These conflicting opinions are illustrated in Figure 2.2.

Statistical challenges

The main challenge with analysis of this data is to estimate treatment effects whilst taking into account the underlying spatio-temporal movements of dingoes as well as the inflated zero problem (Section 2.2). At a site, inflated zeroes occur because the case where there are no visits by dingoes *is confounded* with total absence of dingoes.

Approaches to dealing with zeroes range on a scale between two extremes, and are investigated in more detail in (Pettitt & Low Choy 1999). At one extreme, an ad-hoc approach would discount any blocks from the analysis which had no visits at all. This would preserve balance and remove misleading zeroes from the analysis. Thus the effective number of replications and therefore degrees of freedom for hypotheisis testing are reduced. Estimates of treatment effects could therefore be *over-estimated* if any of the discounted blocks actually contained true zero responses. At the other extreme, all zeroes could be



(a) Expert I: low spatial dependence (b) Expert II: high spatial dependence

Figure 2.2: Differing opinions on spatial behaviour expected of dingos. Chemicals are placed at locations denoted A, B, ..., F.

assumed to be represent true zero responses. Thus the effective number of replications and degrees of freedom are maximized. Estimates of treatment effects could therefore be *under-estimated* if any of the zeroes actually are false zero responses. A naive approach to analysis would fit a GLM to the binary responses and the design matrix, assuming that all zeroes are true zeroes.

At both extremes a generalization had to be applied to all zeroes (within blocks or sites). An intermediate approach would incorporate other information to make a decision on whether each zero represents a true or false zero response. A solution is to model the *conditional* probability of a visit given that a dingo was in the neighbourhood. This necessitates explicitly modelling the presence/absence of a dingo at a pair of locations within a site.

Modelling the underlying presence/absence process, as well as the data model incorporating chemical effectiveness, can be achieved by using a hierarchical model. This challenge was approached following both frequentist and Bayesian paradigms. Once tailored models were constructed, inference had to be especially designed and implemented. (See Chapters 3, 5, 7.) The challenges involved with the Bayesian approach included estimation of a normalization constant (NC) for the underlying spatial presence/absence process (Chapter 6).

Estimating variability of estimated treatment effects provided another challenge. Using asymptotic arguments could lead to results that are not accurate for a small finite lattice such as this one. Additional information on the relationship between sites or blocks were therefore investigated to improve estimates of variability.

2.3.2 Applications to Flora distribution maps: *Cypress* application

A number of **binary two-dimensional lattice datasets**, with **underlying spatial dependence**, have arisen in forestry applications.

Preisler (1993) investigates the spatial process of how bark beetles attack lodgepole pine *Pinus contorta* trees. The primary statistical question was to link tree susceptibility to

covariates specific to individual trees such as age, vigour, size and distances between trees. Of secondary concern was the spatio-temporal spread of the attack. Experimental units were the set of all mature lodgepole trees located in two pure stands of approximately 500 trees and 0.5ha each, situated approximately 25km east of LaPine, Oregon, USA. Observations were made on tree presence/absence and its covariates once each year during the period 1980–1989. The underlying spatial process driving the spread of bark beetles, and therefore the observed clustering of tree absence, was a chemical calling system developed by the bark beetles to concentrate their population at designated targets. Although the trees are not strictly arranged according to a lattice, the concept of immediate neighbours was still well-defined, and distance between neighbours could be incorporated into the covariates.

Another series of studies (Wu & Huffer 1997, Huffer & Wu 1998) devised a model for the distribution of key plant species, depending on climatic variables such as temperature and rainfall. Plant distribution data was digitized from comprehensive atlases of over 180 species. Climatic data was obtained for 9 variables believed to impact on plant distribution. Visual comparison of plant distribution maps clearly showed a tendency for clustering in areas with similar climate, and for neighbouring areas to have similar species present. The presence/absence of plants clearly depended not only on covariates but also on an underlying spatial process.

More recently, Denham & Mengersen (1999) investigate prediction of the presence and absence of white cypress pine in southern central Queensland, based on radar data. These authors are interested in predicting the presence or absence of white cypress pine *Callitris glaucophylla* across a wide area, comprising 2 overlapping areas 75km $\times$ 75km in southern central Queensland. At their disposal they have extremely high resolution radar imagery (approximately 21 million pixels) which has a strong though stochastic relationship with the presence/absence of white cypress pine. Also available is ground-truthing data sampled at 70 sites situated in relatively homogeneous woodland classes. A simplifying assumption is that the ground-truthing data were obtained with no measurement error.

The raw radar data is subject to speckle noise, which occurs when the incoming and reflected radar waves are out of phase resulting in interference in the signal generated. Some methods for filtering the speckle noise in the radar imagery are incorporated into Image Analysis packages such as Imagine (ERDAS 1998). These are generally mathematical morphological techniques such as: back-filtering which do not take into account the stochastic variation or statistical modelling techniques; and pixel-wise maximum likelihood estimation which does not take into account the spatial relationships. Filtering radar data using these standard techniques produces images which show general trends in presence/absence but still suffer from small speckles sprinkled throughout the image. In Denham & Mengersen (1999) both pre-filtered and non-filtered radar data are considered as inputs to the model for presence/absence.

The underlying conceptual model used by Denham & Mengersen (1999) is depicted in Figure 2.3, and to some extent follows that of Wu & Huffer (1997). The observed ground-truthed presence or absence Y depends on whether the grid was searched U and what the true presence/absence Z is at the site. True presence/absence of white cypress pine Z is considered autodependent (neighbouring pixels should have similar values of absence/presence), and in addition depends on an intercept and a covariate for the radar signal X. This places both the errors due to measurement and the errors due to searching at the level of the data model, in $p(y|z,u)$. Spatial variation and error in using SER-1 SAR data for predicting true presence/absence are incorporated at the level of the prior for true presence/absence, in $p(z|\theta,\alpha,x)$, where θ is a vector of spatial parameters, x is a matrix of covariates, and α is

a vector of coefficients for the covariates. This method of accounting for ambiguous zeroes is similar to that of Welsh et al. (1996*a*) given in equation (2.3).

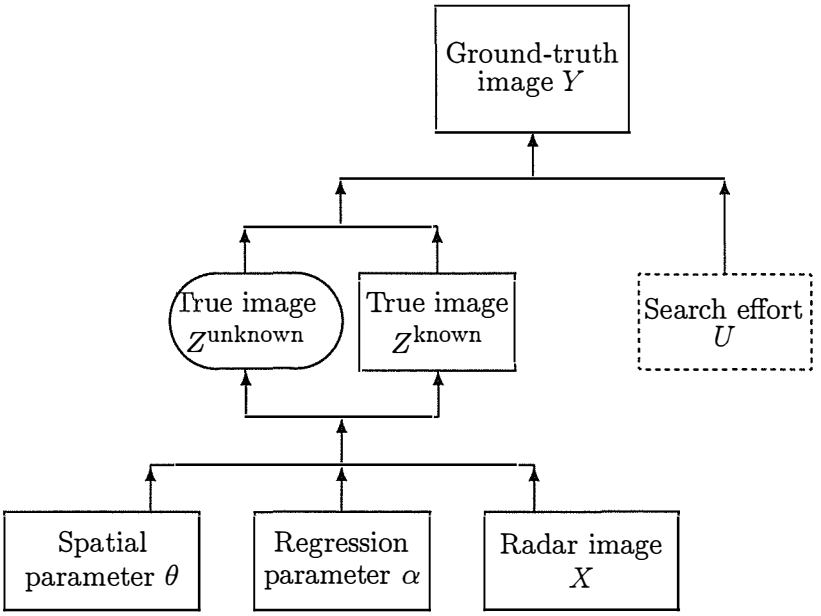


Figure 2.3: Relationship between ground-truthed, radar and true images, and associated parameters as used in Denham & Mengersen, 1999.

Due to the high dimensionality of their dataset (over 21 million pixels in the grid), Denham & Mengersen (1999) limit the hierarchy in the model. They also use pseudo maximum likelihood for estimation by replacing the prior above with the product of all the conditional distributions $\prod_i p(z_i|z_{-i}, \theta, \alpha, x)$. Here i refers to site i and the subscript $-i$ refers to all sites on the grid apart from site i.

An alternative conceptual model can be constructed to provide more separability between errors from different sources, and in particular to isolate spatial dependence from the link between the electronically derived image and the true image. Bayesian hierarchical models used for image reconstruction by other authors (Weir & Pettitt 1997, Aykroyd & Green 1991, Green 1990, Derin & Elliott 1987, Besag 1986), structure the model so that the true image is last in the chain of dependence, *i.e.* the electronically derived image y is obtained from the true image z according to some probability distribution $p(y|z)$. The texture of the true image may then be summarised by some Markov random field model involving components $p(z_i|z_{-i}, \theta)$. Applying this concept here would suggest that the model relationship depicted in Figure 2.4 is a possibility to consider.

Another approach, not considered by these authors, would be to partition the true image Z into two portions, one portion "unknown" and the other "known" by incorporating observed ground-truthed data Y. Then one may assume that the radar data X depends on and is derived from the true image Z. Thus the two main differences between this approach and Figure 2.3 are that: Y becomes subsumed in a partition Z^{known}; and the position of the radar data X is as the dependent variable instead of an explanatory variable for the true image Z. Hence the data model involves $p(x|z, \alpha)$ rather than $p(y|z, u)$ and the prior involves components $p(z_i|z_{-i}, \theta)$ instead of $p(z|z_{-i}, x, \alpha, \theta)$.

Using this alternative formulation it is easy to see the similarity to the *dingo* case study depicted in Figure 5.1. Hence discussion of the *dingo* case study parallels a discussion of this *cypress* dataset.

2.3.3 Biogeographical applications: *Toad* case study

Applications in biogeography are clearly related to the *dingo* and *cypress* case studies, especially where they involve compilation of fauna or flora atlases in the form of presence/absence data on lattices. The output from these analyses are *probability maps* showing probable absence and presence of the species throughout the grid. Natural history texts often contain atlases showing presence/absence of various animal or plant species over large areas such as continents. For example Simpson & Day (1993) give thumbnail sketches of presence/absence for an exhaustive taxonomy of Australian birds.

Recent work in the spatial statistics area (Högmander & Møller 1995) focuses on estimating these distribution maps by applying methods from image analysis for image restoration (Geman & Geman 1984, Besag 1986, Dubes & Jain 1989, for example). Finnish breeding bird atlases are the topic of interest, and extra information is available on the level of research interest which affects the recording of absences. In addition data was accumulated over time, so that analysis could be performed at intermediate stages to ensure sensible evolution of the distribution maps.

The methods of maximum marginal posterior and iterated conditional modes were applied to "restore" the true distribution map z from "noisy" observations y. Maxima or modal values of the posterior distribution of the true map z given y were of interest. Marginal posterior probabilities were also of interest since they can be used to compute the likelihood of breeding in each cell. The computational method used was only partially Bayesian since it was not the full posterior distribution that is obtained, just the maximum or modal values for each component. The authors concluded that the use of incomplete distribution maps can lead to erroneous conclusions.

In Heikkinen & Högmander (1994) the species of interest is the *bufo bufo* toad on a 10km$\times$10km grid superimposed over Finland. Their primary interest was in atlas mapping to estimate the biogeographical range of the toads. Toad presences were reported by a combination of intensive searches by biologists and quality controlled community sightings in some parts of Finland. More effort was expended in areas where toads were thought to be more prevalent *a priori*.

This dataset was found to be particularly challenging due to the high proportion of recorded absence values. The difficulty was that it was impossible to determine whether observed absences reflected true absences, missing values due to lack of searching, or measurement error in accurately recognizing presences. Apparently in these types of situations, visual extrapolation of the true image is often relied upon: atlas maps are often displayed emphasizing only the presences that were recorded; and may be supplemented by maps of search activity. In this case, a wide area of absences recorded in the data obviously correspond to areas of Finland with low human population and therefore low probability of sighting either from biologists or the community.

To supplement the presence/absence data, the authors incorporated information on presence and absence of more common species, the common frog *Rana temporaria* and the common lizard *Lacerta vivipara*, as a proxy for search activity. Instead of a simple count of common frogs or lizards observed in each grid, classes were used to indicate observations of zero, between 1 and 3, or at least 4, frogs or lizards. False absences can then be modelled

within each of the classes, for example by assigning a constant probability to observing an absence when there really is a common toad present. False presences can be ignored in the data due to checking at the time of recording. This approach parallels the use of conditioning on covariates used by Welsh et al. (1996*a*) given in equation (2.3), and the introduction of a latent variable such as that used by Aitchison & Ho (1989) given in equation (2.4).

In this situation, it was of more interest to the biogeographer to estimate the probability of presence for each grid cell, rather than assigning the most likely value of the two values absence or presence. This reflects the various sources of uncertainty in recording absences. In contrast to similar applications in image analysis, the modal overall map was not of interest, rather the modes of each individual grid cell taken from the marginal posterior distributions.

The image restoration method of Högmander & Møller (1995) was unsuccessful due to the large number of uncertain absences. A Bayesian approach to image restoration was used to obtain the complete posterior distribution, rather than just point estimates as for Högmander & Møller (1995). The spatial distribution was approximately estimated by pseudo maximum likelihood. The relationship between variables was also similar in that the observations y depended on the true image z, but in this case also depended on research activity u. The true image in turn depended on a (one-dimensional) spatial dependence parameter θ. In this case the u parameter was also assigned a prior distribution and its posterior distribution estimated from the data and the model. This contrasts with the partial Bayesian approach (Högmander & Møller 1995) where point estimates of the θ parameters were obtained iteratively according to an EM-like approach to estimation.

It is easy to see the correspondence between this situation and that of the *cypress* case study, as illustrated in the graphical representation of the model below in Figure 2.5.

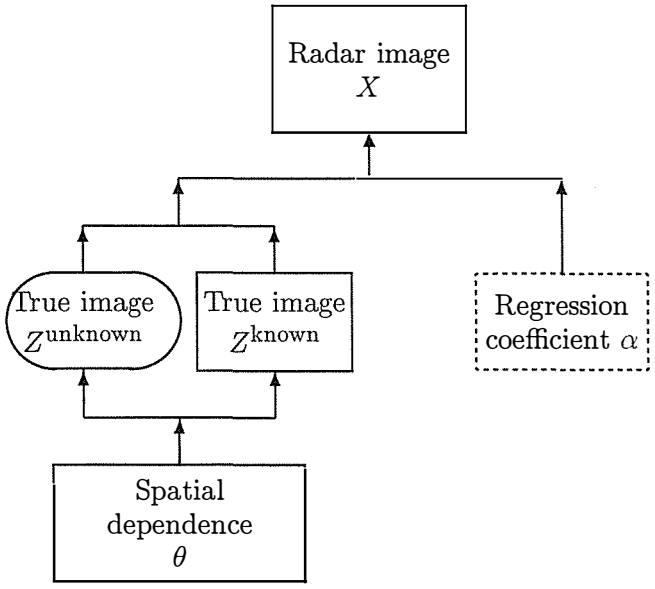


Figure 2.4: Alternative conceptual relationship between ground-truthed, radar and true image variables and other parameters in the *Cypress* case study.

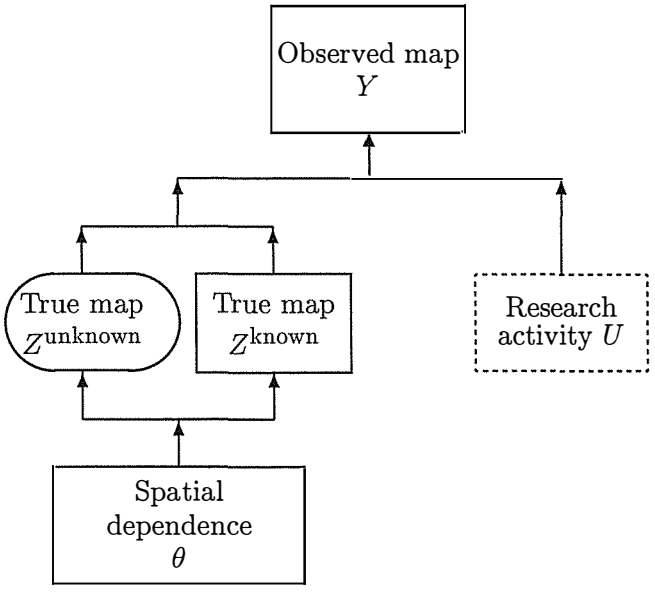


Figure 2.5: Conceptual relationship between observed and true distribution maps, given proxies for research activity and a spatial parameter in the *Toad* case study.

2.4 Overview of models for binary spatial data

Models for binary spatial data have in the main evolved from other models: for multivariate binary data, for time dependent data and for spatially dependent continuous data.

These models can be loosely grouped into the following categories, although there is overlap: multivariate techniques for independent binary data; spatial aggregation of binary data into counts; generalized linear models; Markov random field models; extrapolation of time series models into spatial context; stochastic processes; and kriging. The types of models for binary spatial data are summarized below within each of these categories.

2.4.1 Multivariate techniques

Numerous methods for analyzing multivariate binary data have been applied to spatial data in the image analysis and ecological literature, which due to the extremely large size of datasets requires efficient estimation algorithms. Multivariate methods include discriminant analyis, cluster analysis, canonical correspondence analysis, generalized linear models, and regression trees. For these methods, however, an underlying assumption is independence between individuals. The spatial pattern within the data is only indirectly incorporated via spatial covariates, and not explicitly modelled. Spatial relationships are often visualized by mapping data or model predictions/parameters. A common application is presence/absence data for plant or animal species which can be compared to habitat and climate variables.

Discriminant analysis (DA) was used by Caughley, Short, Grigg & Nix (1987) to determine how 12 range-adjusted climate variables were related to presence/absence of three kangaroo species in Australia. Time effects were assessed by creating variables measuring difference in climate between two time points. Spatial relationships were not capitalized on during analysis. Hjort & Omre (1994) discuss other image analysis applications of DA.

Canonical Correspondence analysis (the binary counterpart of Canonical Correlation analysis) was used by Hill (1991) to relate presence/absence of a group of twenty bird and twenty plant species to a group of environmental variables (representing topography, climate and geology), on a gridded map of Great Britain. Spatial information was only included indirectly via the environmental variables.

Walker (1990) uses classification and regression trees (CART) to analyze similar kangaroo presence/absence data. Regression tree analysis (Breiman, Friedman, Olshen & Stone 1984, Clark & Pregibon 1992) forms decision rules from a training dataset by recursively dividing data into subsets depending on the associations between individuals according to a set of covariates. Threshold values of covariates are chosen to maximize homogeneity in subsets of regression covariates. Spatial information was incorporated indirectly into the model using covariates for latitude and longitude.

2.4.2 Aggregation

A common approach to analysis of binary data relies on aggregation into counts. These counts can then be analyzed by methods based on error distributions such as the Poisson, Gamma or Binomial. For counts aggregated by category, contingency table analysis is often appropriate (Cox 1970, Breslow & Day 1987, Agresti 1990). Generalized linear models (McCullagh & Nelder 1993) can be used to incorporate covariates, with a Poisson error distribution having log or identity link function.

The spatial analogue of these GLMS are the auto-Poisson, auto-Gamma and auto-binomial models of Besag (1974). Explorations into the Poisson and Gamma random fields

have only recently begun within the statistical literature (Wolpert & Ickstadt 1995) since their introduction (Besag 1974). An advantage of the Poisson random field is its infinite divisibility, which allows consistent modelling of counts over a wide range of levels of aggregation, and therefore over a wide range of spatial scales and boundary definitions. This feature is taken advantage of in Wolpert & Ickstadt (1998) to allow modelling and reporting on a range of spatial scales.

Analysis of count data is beyond the scope of this thesis, however, since the motivating examples of interest here all have covariates measured at the level of binary presence/absence, and not at the level of aggregated counts. Some of the methodology described in this thesis may be applicable or extended to count lattice data also.

2.4.3 Generalized Linear models

Where covariates apply to binary presence/absence data, a standard modelling choice in the non-spatial context is a Generalized Linear Model (GLM) (McCullagh & Nelder 1993) with Bernoulli error and logit, probit or complementary log/log link function. Walker (1990) uses logistic regression to model influence of climatic/habitat covariates on presence/absence of kangaroos over a gridded map of Australia. Classification rates were worse than those obtained using regression tree analysis.

Osborne & Tigar (1992) use logistic regression to model presence/absence of three bird species in Lesotho, an African country. The main features of habitat variables indicating land use, geographic location, topography, and vegetation were extracted using Principal Components Analysis (PCA). Classification rates were analysed using jacknife estimates. PCA is used for similar purposes in other contexts: in image analysis; Buckland & Elston (1993) for wildlife distributions. Its advantage is removal of multicollinearity between variables; its disadvantages are that it only considers linear relationships between covariates and interpretation of results may be convoluted.

An omnipresent feature of data collected in plot and field experiments is spatial dependence between observations. Early in the development of statistical science (Bennett 1990), the main strategy for dealing with this spatial dependence was avoidance, using randomization. Experimental designs arrange experimental units in order to randomize, sample or aggregate over spatial effects. The model must reflect the design. Many analytic and computational advantages are obtained from an underlying independence assumption. However extra information is available from the spatial relationships between experimental units, although more sophisticated models are required to make use of this information. This has given rise to various methods for modelling underlying spatial dependence in data.

Progress in spatial experimental design has accompanied the growing sophistication of models incorporating spatial dependence. The first approaches have relied on thoughtful experimental design (see *e.g.* Box, Hunter & Hunter (1978)) to control spatial dependence between observations. Blocks are commonly used to stratify errors within homogeneous spatial regions. The use of split-plots and variations on blocking provide a flexible framework for this stratification. Local spatial effects are effectively aggregated over using these approaches, although low resolution spatial trends can be estimated in this way.

Another modelling innovation which permits handling of spatial or temporal (Stiratelli, Laird & Ware 1984) variation is the use of random effects. The ability to incorporate random and mixed effect models has been facilitated by the development of computational methods, in the frequentist arena, such as Restricted Maximum Likelihood (REML) (Patterson & Thompson 1971) and Minimum Variance Quadratic Estimation (MINQUE) (Rao 1971*a*,

Rao 1971*b*). Standard statistical software has only incorporated these algorithms in the last decade (Payne, Lane, Ainsley, Bicknell, Digby, Harding, Leech, Simpson, Todd, Verrier & White 1987). Bayesian counterparts have also been facilitated by advances in computation, including Gibbs sampling and Markov Chain Monte Carlo (Gelman, Carlin, Stern & Rubin 1995). Again, local spatial effects are either aggregated over to produce estimated effects on the mean response; or else randomized over to produce estimates of components of the variability in the response.

Nearest neighbour models (Bartlett 1978, Baddeley & Møller 1989) were developed to address this need for examination of high resolution or local spatial patterns. One simple nearest neighbour approach is to include a covariate in the linear model which measures an aggregated neighbourhood effect, for example the average or sum of neighbourhood responses. A conceptual drawback of this approach is that a complex and non-hierarchical web of dependence is introduced between the response and covariates. As an exploratory tool this approach is useful provided that careful interpretation is made of results.

One example which led indirectly to this study was investigation of a series of field plot experiments to investigate the spread of disease amongst plants (Low Choy & Pettitt 1992, Chakraborty et al. 1993, Chakraborty et al. 1995). Placing different genotypic lines of the same plant species in close proximity was thought to provide some barrier to the spread of disease. Nearest neighbour methods used to incorporate both high and low resolution spatial variability were based on linear models, with a covariate representing aggregated neighbourhood effects. This stretches the assumption of independence between the tuples of response-covariate values, since each response appears as a covariate in the expression for neighbouring responses.

A nearest neighbour model approach is also taken by Buckland & Elston (1993), where the second stage model predicts future site suitability based on estimates from the first model of suitability at that site and previous presence at that site and neighbouring sites. The manner in which the species moves between sites is also incorporated into a linear model based on log probabilities. Inclusion of neighbouring suitabilities and modelling spatial mobility allows spatial information to be included explicitly in the model. Stochastic models for spatial progression of disease through space are also investigated in other contexts, such as plant disease (Reynolds, Madden & Ellis 1988, Smyth, Chakraborty, Clark & Pettitt 1992, Chakraborty, Pettitt, Cameron, Irwin & Davis 1991).

2.4.4 Markov Random field models

A natural extension of the logistic model (a GLM with logit link function) into the spatial framework is the autologistic model first introduced into a statistical framework by Besag (1974). The autologistic model with three parameters—representing prevalence and spatial dependence separated into North-South and East-West or vertical and horizontal components—is a special case of a binary Markov Random Field (MRF). This approach is suitable for the *Dingo* case study and potentially useful for other biogeographical applications such as the *Toad* and *Cypress* examples. We focus on the autologistic model in more detail in Section 4.3.

Other link functions suitable for binary responses include the probit link and the complementary log/log function. The Hidden Markov Model of Weir & Pettitt (1999) is essentially a probit model based on an unobserved Gaussian variable. This contrasts with the Bernoulli model based on an unobserved Autologistic variable, which is the approach taken in this thesis.

Recent use of binary MRFs as spatial priors within the statistical literature has extended the autologistic model (Besag 1974) to include covariates (Preisler 1993, Huffer & Wu 1998). The theoretical ramifications (Section 4.3.5) of this step have not been addressed in great detail. Instead, in Chapters 5 and 7, I consider a hierarchical model which leaves the pure autologistic distribution intact as an underlying spatial model, and covariates are incorporated into the data.

Zimmerman & Harville (1991) use a random field linear model adopted from geostatistical applications (Cressie 1985, Cook & Pocock 1983, for example) for analysis of field-plot experiments. The model for either point-referenced data or area-referenced data can be written

$$y = X\beta + e$$

where for point-referenced data y is the collection of reponses at each site $\{y_i : i = 1, \ldots, n\}$. The low resolution dependence is modelled via a linear predictor for the mean, where X is the $n \times p$ design matrix; and β is a $p \times 1$ vector of coefficients. The high resolution dependence is modelled by a Gaussian random field e where $\mathrm{E}[e] = 0$ and $\mathrm{Var}[e] = V$ with $v_{ij} = \mathrm{Cov}[s_i, s_j; \theta]$. Here the covariance function depends only on the distance between points s_1, s_2 under the assumption of second-order stationarity. Additional properties of isotropy and separability for the covariance function can improve computations. The authors found that results for their application were robust to the choice of covariance function. Under these assumptions y is multivariate Gaussian with mean $X\beta$ and variance function V. Inference proceeds via Restricted Maximum Likelihood (REML) (Patterson & Thompson 1971), with maximization of the log likelihood implemented by a grid search or golden section search procedure (Kennedy & Gentle 1980, for example) for the examples in that paper. Although not suitable for binary data, this method is related to other methods which are: the Hidden Markov Gaussian model (Weir & Pettitt 1997) which is a hierarchical version; the autologistic regression model (Preisler 1993, Huffer & Wu 1998).

2.4.5 Extrapolation of time series models

Time series analysis is now a well-established area of statistical modelling, including models in the time domain, such as Autoregressive Integrated Moving Average models (ARIMA) (Box & Jenkins 1970), and in the spectral domain (Priestley 1981). In these models covariates enter the model in a linear predictor and residuals capture the time series dependence. The success of these models has prompted extrapolation to the spatial context. Spatial ARIMA and Space-time ARIMA models (Cressie 1993) build on ARIMA models by allowing both spatial and temporal relationships to be modelled after covariate effects have been accounted for:

$$y = \beta^{\mathsf{T}} X + u^{\mathsf{T}} Z$$

where y denotes the response, X is a matrix of known covariates, β is a vector of unknown coefficients, Z is a matrix of spatial relationships, and u is another vector of unknown coefficients. Accurate modelling of the spatial covariance matrix is important, and the spatial variogram is a valuable tool in this process.

Structural time series models using a frequentist or Bayesian approach (West & Harrison 1997) allow the coefficients to vary over time as random walks. To illustrate, a simple

structural model is given by:

$$\begin{aligned} y &= \beta_t x_t + \epsilon_t \\ \beta_t &= \beta_{t-1} + \delta_t \end{aligned}$$

Here the time-indexed covariate x_t enters with time-varying coefficient β_t and an error term ϵ_t. The coefficient β_t then "drifts" over time according to a random walk with parameter δ_t. Distributions are assigned to the error terms (frequentist) or unknown parameters (Bayesian) ϵ and δ. Gaussian distributions are a standard choice:

$$\begin{aligned} \epsilon_t &\sim N(\mu_\epsilon, \sigma^2_\epsilon) \\ \delta_t &\sim N(\mu_\delta, \sigma^2_\delta) \end{aligned} \tag{2.6}$$

Fitting the model in either a Bayesian or frequentist framework proceeds via highly computationally demanding numerical and simulation methods. Allowing coefficients to vary over space using a random walk in two dimensions, has only recently been implemented (Lavine 1999).

Generalized Estimating Equations (GEEs) can be used for analyzing binary longitudinal data (Liang & Zeger 1986, Pendergast, Gange, Newton, Lindstrom, Palta & Fisher 1996) State space models (West, Harrison & Migon 1985) can include covariates, time-varying coefficients and trend parameters. Generalized linear mixed models (Breslow & Clayton 1993) can be adapted for time series applications. Hidden Markov models (MacDonald & Zucchini 1997) are flexible enough to model counts or proportions. See Hay (1999, Chapter 2) for a recent literature review of these methods, and Hay (1999, Chapter 7) for a Bayesian analysis of a time series of counts. The usual approach taken in the GEE methods above is to model the marginal mean as a function of the linear predictor and the marginal variance as a known function of the mean. An alternative is to base the GEEs on conditional means (Zeger & Qaqish 1988).

Albert & MacShane (1995) implement a GEE method for inference of binary spatial data. They have several replicates (one for each patient) of neuroimaging data of the brain which has been spatially referenced on to a common grid. They use a semi-variogram from geostatistics (Matherson 1963, Cressie 1993) to model the spatial covariance matrix. Their model could conceivably be used for modelling the marginal mean probability of presence for situations such as the *Dingo* case study, where interpretation of spatial interactions is with respect to a conditional distribution.

2.4.6 Stochastic processes

Point processes (Ripley 1988) can also be used as models for binary lattice data, where the emphasis is on distance between points and presences tend to be sparse. These models are the close relatives of Markov random fields, but take a subtly different approach to modelling spatial interactions, driven by estimation of a covariance matrix, which depends only on distance between sites. Point processes are unable to capture the onset of cooperative behaviour between groups of points:

> *Phase transitions are essentially cooperative phenomena in which many particles must interact with each other at the same time. Dilute systems in which particles are, on the whole, independent of each other, do not exhibit phase transitions. (Domb & Green 1972a, preface)*

Longitudinal data can be treated as "repeated measures" within a linear model context. The usual linear model methods (Box et al. 1978) may be used for accounting for covariates and the experimental design. Correlations due to repeated binary observations over time measured on individual units may be modelled with a single parameter. (Cox 1970). Spatial correlations between individual units have also been accounted for in similar ways. Definition of a block structure can describe a spatial structure in a discretized fashion. Differences between blocks can then be fitted using an explanatory factor; trends over blocks may be modelled using a covariate at block level to indicate spatial continuity. Green, Jennison & Seheult (1985) takes another approach whereby residuals are spatially smoothed using least squares. In Pettitt & Low Choy (1999) included in Chapter 3, both the block and spatially smoothed residual approaches are investigated.

Random and mixed effects in hierarchical (multi-stratum) linear models are also used to handle spatial variability. Stiratelli et al. (1984) suggest a random effects model for binary data. Two models are devised by Buckland & Elston (1993) to examine suitability (probability of presence) of two bird species and one deer species. Absence/presence data is collected *at random* locations over the grid to preserve independence assumptions. The first model for suitability allows for covariates and error structures at two levels of spatial resolution. Suitability is first modelled at the low level of resolution using low resolution covariates combined with aggregations (means) of high resolution covariates and random effects within lower resolution areas. Then suitability is predicted at the high level of resolution using low resolution covariates and disaggregated high resolution covariates. Spatial information is used to estimate random effects within and between low resolution sites, similar to the block varying estimates of presence described above.

2.4.7 Geostatistical models

Kriging (Matherson 1963, Cressie 1993) is another method of modelling spatial variation of a response. Here the emphasis is on interpolating (predicting) a surface given point estimates at various locations. Analysis relies on modelling the spatial variogram (the spatial analogue of the auto-correlation function in time series analysis). Assumptions of second order stationarity in this variogram, imposing structure on the low resolution spatial process, facilitate prediction. High resolution or local spatial variation is estimated by components of the variogram corresponding to short distances. However these methods are best developed for continuous variates. Indicator kriging can be applied to binary data, although an underlying unobserved continuous variate is assumed. It is possible to incorporate effects of covariates using a method termed 'co-kriging'. This approach is not appropriate here since the main aim is not prediction of responses, but in explanatory modelling of effects of covariates.

2.5 Discussion

Thus there are some immediate applications for this research, in particular

- field experiments with underlying spatial variation such as the *dingo* case study,
- forestry applications such as the *cypress* case study,
- and biogeographical population atlas applications such as the *toad* case study.

A particular modelling problem common to all these fields of study is that of estimating effects of covariates, whilst accounting for underlying spatial variation.

Currently available methods can be applied to the task, but have various drawbacks. Multivariate methods are limited in general to be non-hierarchical, and therefore do not permit the flexibility of modelling separable errors arising from different sources and processes. However they provide preliminary methods for selection of important covariates. Various enhancements to generalized linear models are available to cope with underlying spatial variation, including block and hierarchical error structures, random and mixed effects model and nearest neighbour models. These models however are not capable of modelling local spatial relationships whilst maintaining consistency in the global relationships. Other modelling approaches—such as aggregation of data into counts; geostatistical models; point processes; spatial extensions of time series models—do not address the research question of interest.

This research fills a niche in the established literature for modelling binary two-dimensional lattice data. Hierarchical models, in either a frequentist or Bayesian context, are alternatives we will explore further in this thesis. The hierarchy should allow for a data model which explains dependence on covariates, as well as a spatial dependence model to explain high resolution (local) spatial correlation between responses and/or covariates. The framework already provided by generalized linear models can be easily applied to the data model where responses are binary. Markov random fields appear to be promising candidates for the underlying spatial dependence models.

Chapter 3

Frequentist hierarical models for 2D binary lattice data with underlying spatial dependence

Contents

3.1 Introduction **33**
3.2 Exploratory Data Analysis for *Dingo* case study **33**
3.2.1 Raw data 33
3.2.2 Exploratory data analysis 35
3.2.3 Sparseness of visits 37
3.3 Frequentist hierarchical modelling: transcript of published paper **40**
3.3.1 Introduction 42
3.3.2 The Design and Data 43
Dingo behaviour 43
Design of field experiment 44
The data 45
3.3.3 Analyses for chemical effectiveness 46
Initial Analysis 46
Conditional Analysis 47
Full data model involving dingo presence 48
3.3.4 Results of the basic modeL 50
Estimates and standard errors 50
Discussion 51
3.3.5 Bootstrap estimates incorporating time dependence 51
3.3.6 Further remarks 54
Modelling spatial smoothness 54
Analyses using Bayesian Techniques 55
Conclusion 55
3.3.7 Appendix A 58
Information for unconditional/conditional analysis 58
3.3.8 Appendix B 60

Standard Errors for EM Algorithm Estimates 60
3.4 Discussion 62

3.1 Introduction

In the long run we are all dead.
- John Maynard Keynes, "A tract on monetary reform" (1923)

In this chapter the *dingo* case study first introduced in Section 2.3.1 is explored more fully in a data analytic and frequentist context. Section 3.2 introduces analysis of the *dingo* experiment with an exploratory data analysis.

Section 3.3 presents work published with the supervisor (Pettitt & Low Choy 1999). This paper gives a full explanation of the dingo experiment, presents a hierarchical model used to cope with ambiguous zeroes, and explains the frequentist approach taken to analysis. The EM algorithm was used to estimate parameters, and both asymptotic and bootstrapped standard errors were computed.

Later within the thesis, Section 4.5.1 covers issues on parameterization which apply both to the frequentist analysis (Section 3.3) and Bayesian analyses (Chapters 5–7.)

Appendix B presents additional work on standard error computations using the bootstrap, that was not included in the paper.

Finally Section 3.4 summarizes this chapter. Conclusions presented in the paper are not re-iterated in these conclusions.

3.2 Exploratory Data Analysis for *Dingo* case study

The biological, environmental management and statistical questions relating to the *dingo* case study were outlined in Section 2.3.1. In this section I present a brief summary of the observations collected from the experiment since exploratory data analysis should precede more formal analysis.

The raw data are displayed in a series of graphs in Section 3.2.1. The binary responses can be aggregated (Section 2.4.2) into counts to summarize overall patterns. Contingency tables of counts, by chemical type, by day, and by both illustrate the apparent attractiveness of chemicals, and the variability over days. The sparseness of dingo visits is also discussed.

3.2.1 Raw data

An image displaying the observed visits to pairs of chemicals at sites is shown in figure 3.1. Vertical stripes highlight strings of visits to the same site along the transect on different days. Dingoes repeatedly visited sites 8, 41, 100, 111, 114. Horizontal stripes highlight strings of visits on the same day to neighbouring sites along the transect. There were only 12 instances where a pair of neighbouring sites were visited on the same day, and 5 instances for a triplet of neighbouring sites. Visually, there appear to be many visits to sites which are not accompanied by visits to neighbouring sites in time or space. There also appear to be oblique striations[1] in the data, particularly in the first 20 sites along the transect, from sites 60 to 100, and the last 20 sites. This might arise from dingoes systematically exploring just a little further along the transect on consecutive days.

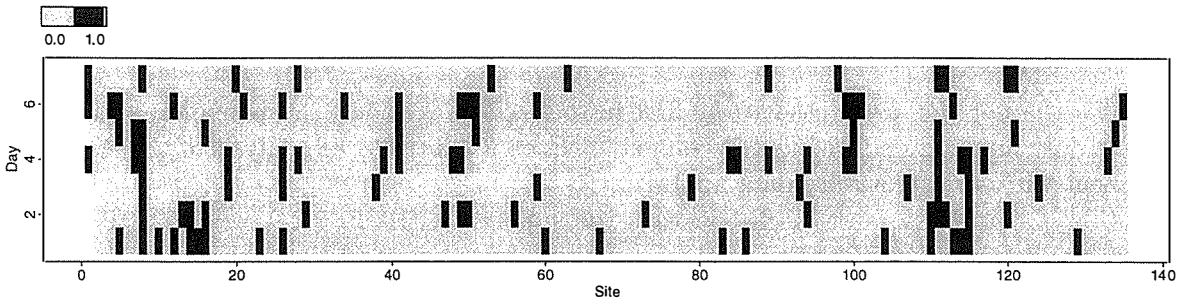


Figure 3.1: Observations: positions where a dingo visited at least one of the locations ($y_{.st} > 0$), so was definitely present ($z_{st} = 1$) are coloured black. This notation is explained in Table 3.5. The x-axis represents spatial location (site along the transect) and the y-axis temporal location (day).

Table 3.1: Frequency of dingo presence at a site with designated pair of chemicals, classified by chemical pair (row) and by day (column). Note one presence is counted if there is at least one visit to either chemical in the pair.

	Day							
Chemical Pair	1	2	3	4	5	6	7	Total
AB	2	0	1	2	0	4	0	9
AC	0	0	1	1	1	0	1	4
AD	3	3	1	1	2	1	1	12
AE	3	2	2	2	1	0	2	12
AF	2	1	2	4	2	2	2	15
BC	0	1	0	0	0	2	0	3
BD	1	1	0	0	1	3	1	7
BE	2	2	1	1	1	0	1	8
BF	0	0	0	0	0	0	0	0
CD	0	1	1	3	1	2	1	9
CE	1	2	0	2	0	0	0	5
CF	1	1	0	0	0	0	1	3
DE	1	0	1	1	1	1	0	5
DF	2	1	0	1	0	0	2	6
EF	1	1	1	3	0	2	0	8
Total	19	16	11	21	10	17	12	106

Possible total number of presences in: each cell (day-chemical combination) is 9; each day is 135; each chemical pair is 63; and overall is 945.

Table 3.2: Frequency of dingo visits to locations at a site with designated chemical, classified by chemical (row) and by day (column). Note that a visit is counted for every location visited which has that chemical.

	Day							
Chemical Pair	1	2	3	4	5	6	7	Total
A	8	6	3	9	3	5	4	38
B	1	1	1	0	0	2	0	5
C	1	2	2	0	1	2	1	9
D	7	3	2	6	3	6	4	31
E	7	5	4	6	2	1	1	26
F	3	3	2	3	2	3	3	19
Total	27	20	14	24	11	19	13	128

Possible total number of visits in: each cell (day-chemical combination) is 45; each day is 270; each chemical is 315; and overall is 1890.

3.2.2 Exploratory data analysis

Table 3.2 summarises the number of locations visited by dingoes, classified by day and chemical. The total possible number of visits to locations is $135 \times 2 \times 7 = 1890$. Table 3.1 shows the number of visits to each pair of chemicals, tabulated by day and chemical pair. A visit to at least one chemical in the pair scores 1, and visit to neither chemical scores 0. The total possible number of visits to sites is $135 \times 7 = 945$. Note denominators for counting visits to each chemical differ from those for each chemical pair.

Chemicals A, D, and E were very well visited with 38, 31 and 26 total visits out of a possible 128. Chemicals B and C were not so popular, with 5 and 9 visits respectively. Day 1 had the most number of visits (27), whereas days 5, 7 and 3 had the least number of visits (11, 13, 14). Visits to locations could not be explained, using a GLM with logit link for binomial data, by the main effects of factors representing chemical and day alone ($R^2 = 6.23\%$). In this context R^2 represents the change in deviance obtained by fitting this model in comparison to a null model. These effects were significant, but explained very little of the variation in the data.

The chemical pairs with the most visits all involved chemical A—the other chemicals were D, E, and F; with 12, 12, and 15 visits respectively. Chemical pairs with a moderate number of visits were chemicals AB, CD, (both 9 visits) and BE and EF (both 8 visits). Chemical pairs with the least visits mostly included chemicals B, C, and F: AC (4 visits); BC and CF (3 visits); and BF (no visits). It is surprising given the individual performance of chemicals D and E that the combination of the two did not work well together. This indicates that something other than the main effects of chemicals are at work here, for instance interactions or a latent variable. Visits to sites could not be explained, using a GLM with logit link for binomial data, by the main effects of factors representing chemical pair and day alone ($R^2 = 7.41\%$), where R^2 represents the change of deviance in this context. These effects were significant, but did not explain all the variation in the data.

Table 3.3 shows the most popular and least popular chemical pairs, by day. Daily visits to the most popular chemicals A and D had a different pattern to E which was not visited

[1] **stria** *n.* a linear mark, slight ridge, or groove on a surface, often one of a number of similar parallel features. **striated** (technical) *adj.* marked with stria. *derivation* **striation** *n.* (Pearcall 1998).

as much on Day 6. The daily pattern of visits to popular chemicals also differed from B, C and F whose visits appeared almost uniformly distributed over days. The exception was for chemicals B, C, and F, there was a surprising lack of visits on Day 4 which was the most, or second most, popular day for the other chemicals A, D, and E.

Table 3.3: Ranking of visits to chemical pair, by day. Superscripts indicate where visits exceed 2.

Day	Most popular chemical pair (4,3,2 visits)									
1	AB	AD^3	AE^3	AF		BE			DF	
2		AD^3	AE			BE		CE		
3			AE	AF						
4	AB		AE	AF^4			CD^3	CE		EF^3
5		AD		AF						
6	AB^4			AF	BC	BD^3	CD			EF
7			AE	AF					DF	

Day	Least popular chemical pair (0 visits)										
1		AC		BC		BF	CD				
2	AB	AC				BF			DE		
3					BD	BF		CE	CF	DF	
4				BC	BD	BF			CF		
5	AB			BC		BF		CE	CF	DE	DF
6		AC	AE		BE	BF		CE	CF	DF	
7	AB			BC		BF		CE		DE	DF	EF

Table 3.4 shows both the relative relationships between chemicals as well as the absolute attraction of each. It is apparent that every day, the most visited chemical was A or D or E. Virtually every day chemical A was visited. Exceptions were Day 4 when E had one more visit than A, and Day 6 when D had one more visit than A. Every day, the least visited chemical was one of B, C or E. Virtually every day chemical B was least visited. The exception was Day 6 when chemical E had one less visit than B. Chemical F consistently ranks in the middle order between positions 3 and 4.

It is also interesting that the chemicals with a high number of visits on any given day (A, D, E) had the most variable number of visits. Chemical A had from 9 down to 3 daily

Table 3.4: Ranking of visits to chemicals, by day

	Number of visits to chemical									
Day	9	8	7	6	5	4	3	2	1	0
1		A	D,E				F		B,C	
2				A	E		D,F	C	B	
3						E	A	C,D,F	B	
4	A			D,E			F			B,C
5							A,D	E,F,F	C	B
6				D	A		F	B,C	E	
7						A,D	F		C,E	B

visits; Chemical D had from 7 down to 2 daily visits; and Chemical E had from 7 down to 1 daily visits. Chemicals B and C had 0, 1, or 2 daily visits. Chemical F consistently had 2 or 3 daily visits.

3.2.3 Sparseness of visits

On closer inspection sparse visitations by dingos on some days could be affecting these simple totals quite drastically. See Figure 3.2 for an illustration of the number of visits to sites (0, 1, or 2) on the y-axis, against sites on the x-axis, repeated for every day allocated a separate panel on the y-axis. For instance on day 5 ten consecutive blocks (of three sites, each with two locations) were not visited at all by dingos.

Figure 3.3(top) aggregates these figures over days, to show the total number of visits to sites (ranging between 0 and 7) on the y-axis for each site shown on the x-axis. There appear to be some clusters of sites having more visits than others, near sites $0-30$, $45-60$, and $90-125$. Sites between 61 and 89 are obviously less visited than other sites. The lower portion of the figure shows a similar relationship for the number of presences rather than visits to sites. A similar pattern emerges.

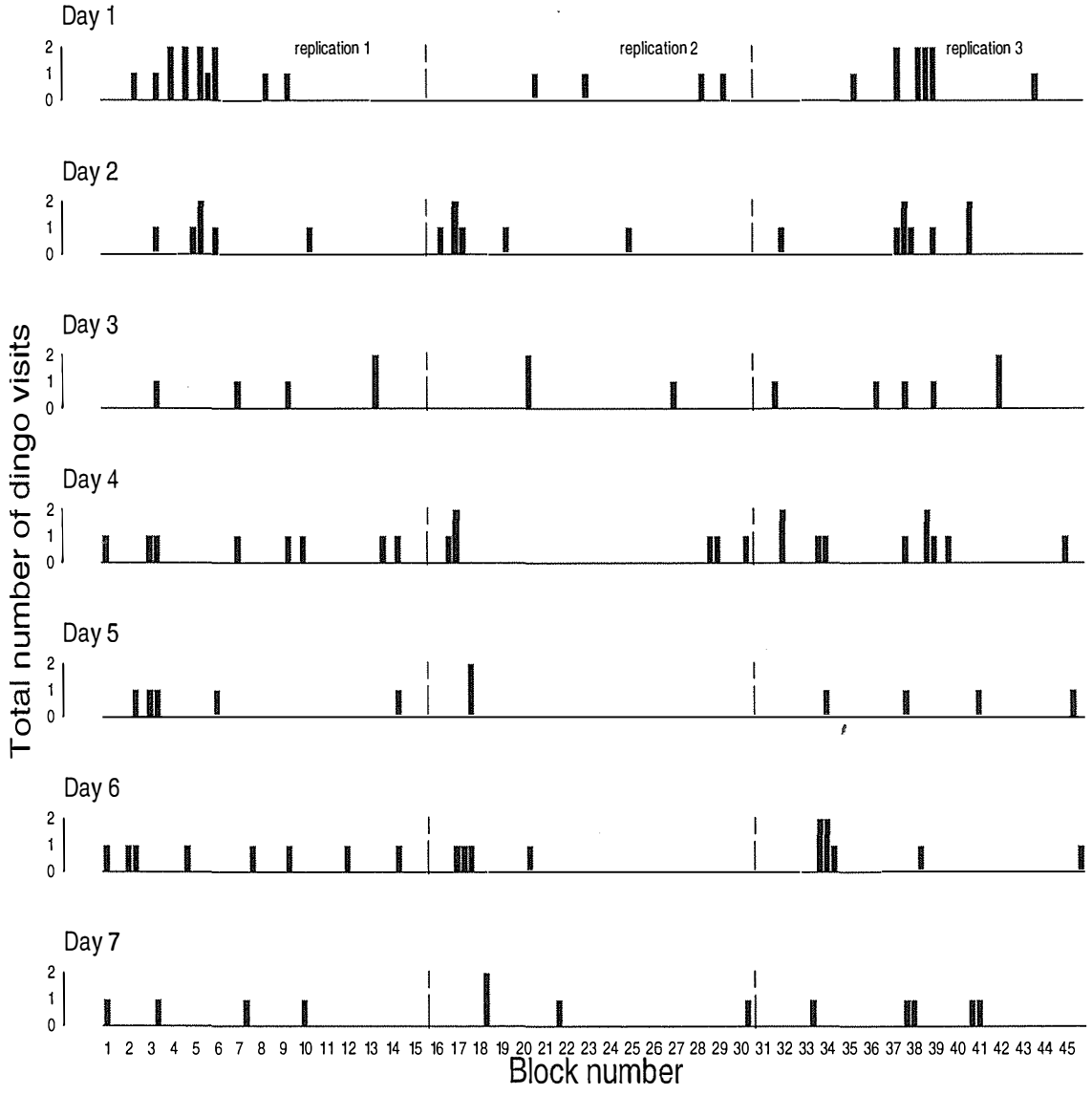


Figure 3.2: Total dingo visits, to each site for each day in the experiment. Three sites per block.

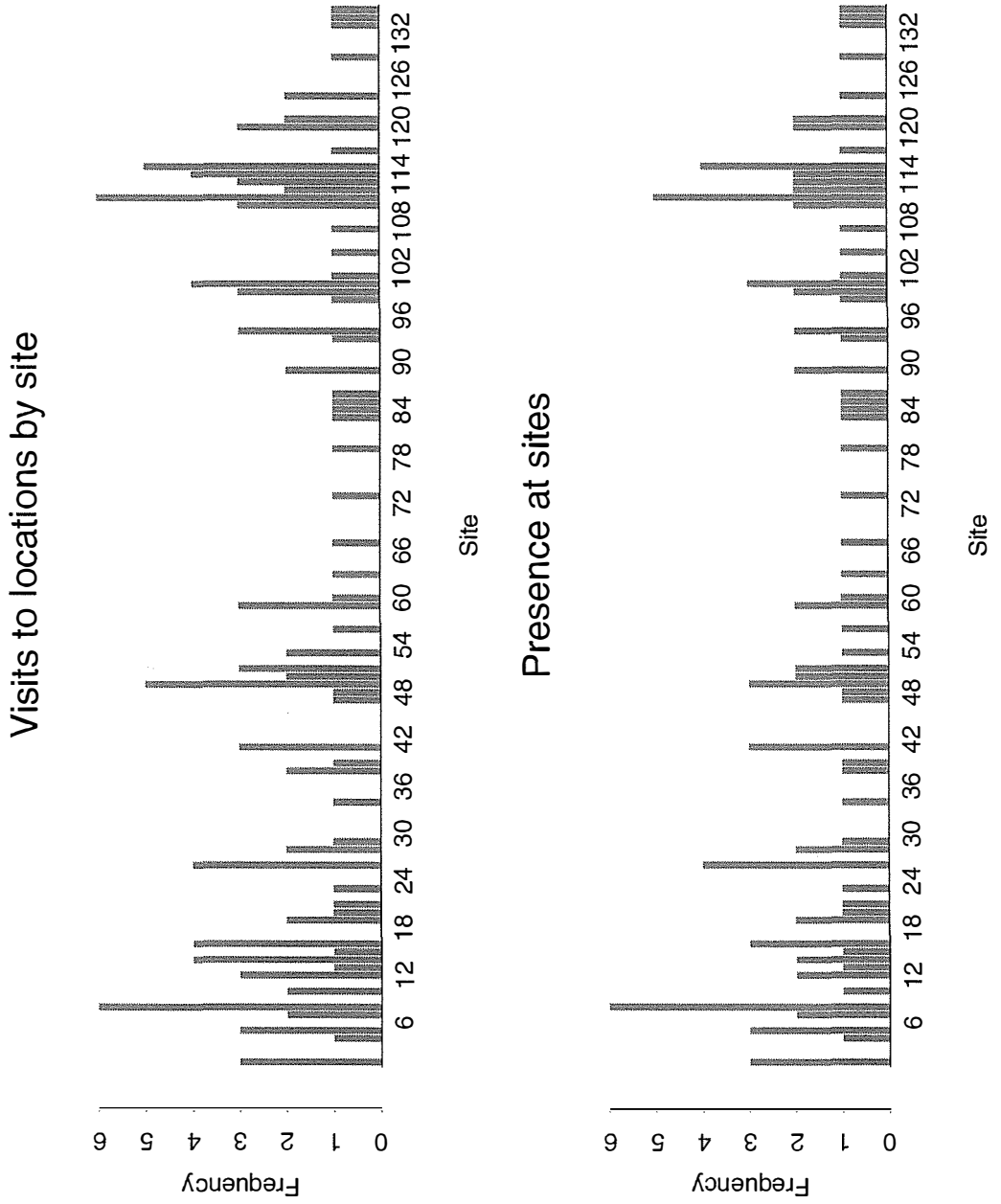


Figure 3.3: Total dingo visits or presences at each site. Three sites per block.

3.3 Frequentist hierarchical modelling: transcript of published paper

Introduction to published paper

This section focuses on the illustrative *Dingo* case study. It presents work published as Pettitt & Low Choy (1999). This work pursues the questions put forward in Section 2.3.1, and expands on the exploratory data analysis presented in Section 3.2.

In this work, we use a hierarchical model (similar in some ways to that used in Chapter 5) and a frequentist approach to the analysis of the *Dingo* case study. The data model is a Bernoulli distribution for dingo visits depending on chemicals applied at locations, as well as dingo presence. This same data model is used for the Bayesian approaches in Chapters 5–7.

Two different methods are used for modelling underlying dingo presence: block probabilities of dingo presence and probabilities smoothed between adjacent sites. These methods contrast with the autologistic model used in the Bayesian context which simultaneously models presence at positions in the lattice, as depending on neighbouring positions. Two different methods for estimating standard errors are used: asymptotic and bootstrapped.

The notation employed in the paper is preserved here and is specific to the application. It differs from the more general notation used in this thesis. In Table 3.5 is a correspondence table of notation to assist in translating between the paper and the rest of the thesis.

Table 3.5: Equivalence between notation used in paper Pettitt & Low Choy (1999) and the rest of the thesis.

Pettitt & Low Choy (1999)		Thesis	
Notation	Meaning	Notation	Meaning
i	site	s	spatial position (can be multi-dimensional)
j	location	v	variable
t	day	r	time period
		t	Monte Carlo simulation iteration
		i	spatio-temporal position (can be multi-dimensional)
k	chemical index	k	treatment index
τ	chemical indicator variable	τ_i	treatment indicator variable
y_{ijt}	visit to site i, location j, day t	y_{vi}	presence/absence variable v observed at spatio-temporal position i
D_{it}	presence at site i, day t	z_i	underlying spatio-temporal dependence process z at position i
q_k	probability of dingo visit to chemical k given dingo present	q_{vi}	probability of success with treatment allocated to variable i, position i
		α_k	inverse logit probability of success of treatment k given presence

Bivariate binary data with missing values: analysis of a field experiment to investigate chemical attractants of wild dogs[2]

A N Pettitt and S Low Choy
School of Mathematical Sciences
Queensland University of Technology
G P O Box 2434
Brisbane 4001
Queensland
Australia

Summary

This paper considers design of a field experiment to investigate the effectiveness of various chemical lures or attractants for dingoes, *Canis lupus dingo*, Australia's native wild dog, and the subsequent analysis of the resulting data. Chemicals were located 50m apart at each of 135 sites equally spaced at 500m apart along an approximate straight path about 70km long, a so-called transect design. Successful attraction to the chemicals was noted each day for seven days. Analysis of the resulting bivariate binary data (successful or not) are carried out to obtain estimates of the effectiveness of the chemicals. Where there was no response at a site this could either have been due to failure of the chemical to attract a dingo, which was present, or absence of the dingo from the site. In order to analyse the resulting data, a model conditioning on dingo presence/absence and hypothesising a distribution for dingo presence/absence is introduced and estimates of the attractiveness of chemicals (defined as a probability) are obtained using an EM algorithm. Standard errors of estimates are obtained using both asymptotic approximations and a bootstrap for dependent data. An analysis which conditions on observing at least one chemical at a site being visited by a dingo is investigated and estimates obtained. An investigation is made of the information lost by the conditional analysis. Both empirical and theoretical results infer precision is gained by considering the unconditional analyses.

3.3.1 Introduction

This paper considers the analysis of data from a field trial carried out to investigate the effectiveness of various chemical attractants for Australian dingoes (*Canis lupis dingo*). The field trial involved placing samples of the six chemicals along a transect. The "transect" was defined by following earth tracks and roads, fence lines etc., in an approproximate straight line for about 70km. Only sparse visits of dingoes to sites occurred with about 15% being visited daily over the seven days. Given dingoes known behaviour, it is quite likely that for certain parts of the transect on certain days dingoes are not present and a non-response is certain irrespective of the chemicals' attractiveness. We therefore develop an analysis that explicitly obtains information on the probability of attractiveness of a chemical given a

[2]**Keywords:** animal presence; binary data; bootstrap; *Canis lupus dingo*; dependent data; EM algorithm; missing data; transect design.

dingo's or dingoes' presence. In order to do this, possibly missing variables are introduced for a site where the outcome is such that both chemicals fail to attract dingoes, because, for such observations, one cannot distinguish between dingo absence at the site and the dingo not being attracted given dingo presence. We develop an EM algorithm to find maximum likelihood estimates of chemical effects. The model requires the probability of dingo presence to be constant over regions of the transect through time and we investigate the sensitivity of the estimates to the choice of region. Standard errors for the estimates are found using both standard asymptotic theory and a dependent data bootstrap which takes into account the possible time dependence over days at a site due to the same chemicals being located at that site over the seven days of the experiment. A conditional analysis, which does not require the dingo presence variables to be introduced, is also considered. However, the conditional distribution does not contain all the information about the chemical effects and this loss of information is investigated.

In section 3.3.2 we give some background on dingoes, the design and the data; in section 3.3.3 the model is developed and an EM algorithm given. Section 3.3.4 gives results and Section 3.3.5 considers bootstrap estimates. Section 3.3.6 considers a smoothed version of the EM algorithm, which appears to have poor convergence properties, and contains further remarks and suggests an alternative Bayesian analysis. Appendices consider the efficiency of the conditional analysis and large sample standard errors.

3.3.2 The Design and Data

Dingo behaviour

Corbett (1995, Chapter 4) gives an extensive account of the spatial organisation of dingoes in Australia and some relevant features follow. Dingoes can either belong to packs with well defined territories or be 'loners'. Such territories or home ranges can vary in size from 10km to 80km depending upon habitat. Ranges of different packs can overlap and separation of packs and dingo individuals is essentially spatial but when resources (*e.g.* water) are shared separation tends to be temporal. Typically, there would be as many dingoes in packs as individual loners and average pack size can vary from 5 to 12 individuals.

Daily distances travelled by dingoes depend upon habitat but typical distances are 10 to 20km per day. Activity carries on throughout the 24 hour day with short periods of rest.

Radio tracking of animals suggested two types of movement patterns. *Exploratory* movement was characterised by apparently purposeful movement from one place to another and the traversal of a substantial area. At the end of exploratory movement, dingoes tended to follow *searching* movement, characterised by intensive activity in a small area and frequent changes of direction. Thus the movement of the dingoes along the experimental transect (or approximately straight path of about 70km length) is difficult to postulate without detailed local knowledge, but it is plausible that dingoes would move distances of the order of 5–10km along the transect in a day and make intensive searches over a small area of 100m or less.

Dingoes use faeces/urine scent-posts for *inter alia* marking territory, as signs to hunting grounds and for mating. Thus to attract dingoes, substances resembling the scent of faeces or urine should be successful. However, in an earlier Queensland study, Mitchell (1988) suggested that the attractiveness of a lure could depend upon the density of dingoes. In areas of high density, dingoes would tend to be attracted by smells with behavioural or social importance, whereas in low density social stress would be replaced by food stress.

Attractiveness might also vary with season, so in the mating season appropriate lures would be related to sexual stimuli while, in the pup season, young dingoes would be attracted by different smells from those for adults.

Design of field experiment

In an earlier experiment, Mitchell (1988), similar experimental transects 200km long were used and single chemicals placed at 500m apart offset from the transect by 50m, alternating from side-to-side. Eight chemicals were investigated and a "block" defined by randomizing the order of these chemicals at 8 adjacent positions on the transect. Blocks were replicated and separated by 3- 5km along the transect. In this earlier experiment, the best chemical had a daily visitation rate equal to 14%. Given the likely behaviour of dingoes, the purpose of the new design was to take advantage of dingoes searching behaviour more effectively by placing two attractants at a site 50m apart at right angles to the transect, and sites 500m apart all along the transect.

The final design involved a block which consisted of three sites arranged 500m apart along the transect. The six chemicals were arranged in pairs, one pair at each of the three sites. If the within pair arrangement and the order of the three pairs within a block is ignored, there are fifteen distinct ways of choosing the three pairs of chemicals. If fifteen blocks are arranged in this way then each chemical is paired at a site with any other chemical in three out of the fifteen blocks. These fifteen blocks were then replicated three times along the transect giving 135 sites each 500m apart, with the two chemicals within a site located 50m apart at right angles to the transect. Within the set of fifteen blocks, the order of blocks was randomised and, within blocks, sites for pairs of chemicals were also randomised. Part of the design is illustrated in Figure 3.4 and the full design is available from the web site at the end of the paper.

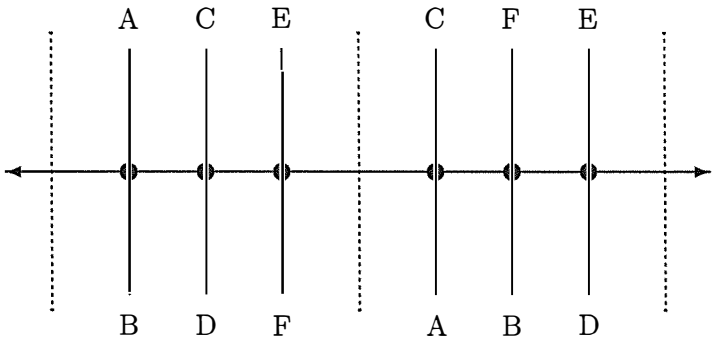


Figure 3.4: Illustration of two blocks from the transect design. Horizontal and vertical distances are not to scale. Chemicals $A - F$ were placed in pairs within blocks of six.

We note that if the design had had only one chemical at each site rather than two, then chemical attractiveness and dingo presence at the site would have been confounded. With three or more chemicals at each site there might be considerable interaction between the effectiveness of chemicals. Additionally, earlier pen trials were carried out with close proximity of different chemicals and the field experiment was designed to provide different information.

In the experiment, the sandy soil surrounding each location where a chemical was positioned was smoothed each day after first noting whether a dingo or dingoes had been

attracted to within a few metres of the chemical, as indicated by paw prints and other signs. The same chemical was placed at a given location on each day of the experiment, because of (i) the practical difficulty of adhering to a more complex design which changed chemicals from day-to-day, and (ii) the probable carry-over effect if different chemicals were used at the same location. Of course, some sites might be intrinsically more attractive to dingoes but randomisation and balance in the design minimise such an effect for comparisons amongst chemicals. We note we use site to refer to the area about 50m across where the two chemicals were situated, and location to the place where each chemical was placed.

If the response had been a continuous value at each location with no missing data, the difference of the measurements for the two locations for a given site could be analysed to investigate the effects of the chemicals. These location differences for each site could then be considered to be independent over sites within a day. Standard analyses carried out take into account the block, replication structure and the repeated measures (over days) nature of the data, *i.e.* the responses on different days at the same site might be dependent. However, for this experiment, the realised data were somewhat different from this assumption and new methods were demanded.

It has been suggested that a "control" could be placed at a location to determine dingo presence/absence. In fact, one of the six chemicals was water and this begs the questions (i) what is an appropriate control, (ii) is the control placed once at each location, site or block, and so on. The design used had the "control", water, placed once in each block. It is not certain that a dingo would always be attracted by the control and so presence/absence would become known.

The data

The anticipated sparsity of dingo visits occurred in the experiment with about 15% of sites being visited each day. The summary information in Table 3.6 indicates how successful the various chemicals were. An initial analysis might treat the number of successful occasions when dingoes were attracted as being binomial, but there is obvious dependence, certainly over time, which makes this untenable. Also, such an analysis is not straightforward because the total number of sites where dingoes were present is unknown. The total absence of a dingo from a site obviously gives the same response as non-attractance to both chemicals at the site. This problem of "extra-zeroes" which might be expected from the sampling distribution is common in abundance studies, *e.g.* Welsh (1996). Alternatively, one might only consider those sites where at least one chemical attracted a dingo, that is, condition on those sites where it is known for certain that a dingo was present.

In Figure 3.5 we give an "image" of known dingo presence along the transect over the seven days of the experiment. Looking at this space-by-time view of dingo behaviour in terms of presence/absence it is obvious that dingoes appear to have been absent from considerable parts of the transect for all of the time. On an *ad-hoc* basis we could divide the space-by-time array into regions of supposed presence/absence and analyse the responses from the 'presence' region. Therefore, at one extreme of this approach, we would assume that dingoes were present at all sites and, at the other, we would assume that dingoes were present only at those sites where they were known to be present, that is, at least one chemical at the site attracted dingoes. The total number of trials in which a chemical was assumed to be tested could therefore vary from a small percentage (typically 15 to 20%) of the total number of location-site-day combinations to the total. We also note by considering north-south and east- west clusters of two and more pixels in the image that temporal and

Table 3.6: Number of chemically treated lures to which dingoes were attracted, classified by day and chemical

Chemical	Day							Total
	1	2	3	4	5	6	7	
A	8	6	3	9	3	5	4	38
B	1	1	1	0	0	2	0	5
C	1	2	2	0	1	2	1	9
D	7	3	2	6	3	6	4	31
E	7	5	4	6	2	1	1	26
F	3	3	2	3	2	3	3	19
TOTAL	27	20	14	24	11	19	13	128

spatial clustering appear about the same. In the next section we present an analysis which models the dingo presence explicitly.

Table 3.7: Chemical effect estimates for varying number of blocks over which probability of dingo presence is constant using ML

No. sites over which p_D is constant	q_A	q_B	q_C	q_D	q_E	q_F	No. p_D parameters*
$N_s = 3$	0.608	0.085	0.162	0.537	0.436	0.331	315
$N_S = 5$	0.548	0.085	0.160	0.525	0.418	0.328	189
$N_S = 9$	0.555	0.076	0.151	0.497	0.413	0.323	105
$N_S = 15$	0.541	0.077	0.150	0.489	0.407	0.314	63
$N_S = 135$	0.522	0.079	0.153	0.505	0.409	0.307	7

* does not take into account cases for which no information for given p_D is available.

3.3.3 Analyses for chemical effectiveness

Initial Analysis

We first take an extreme approach which assumes dingoes were present at each site. We consider the estimates of chemical attractiveness which give the proportions of occasions when a chemical was successful. We obtain estimates for $q_A^u, \ldots, \ldots, q_F^u$, the unconditional probabilities of attraction for the chemicals $A, B, \ldots, F$.

$$\begin{aligned} \widehat{q}_A^u &= 38/315 = 0.121, & \widehat{q}_B^u &= 5/315 = 0.016, \\ \widehat{q}_C^u &= 9/315 = 0.029, & \widehat{q}_D^u &= 31/315 = 0.098, \\ \widehat{q}_E^u &= 38/315 = 0.121, & \widehat{q}_F^u &= 5/315 = 0.016. \end{aligned} \tag{3.1}$$

These probabilities reflect that only about 15% of sites were visited. Many "zeroes" are due to absence of dingoes rather than lack of attractiveness of the chemical. These data do, however, give information about relative effectiveness.

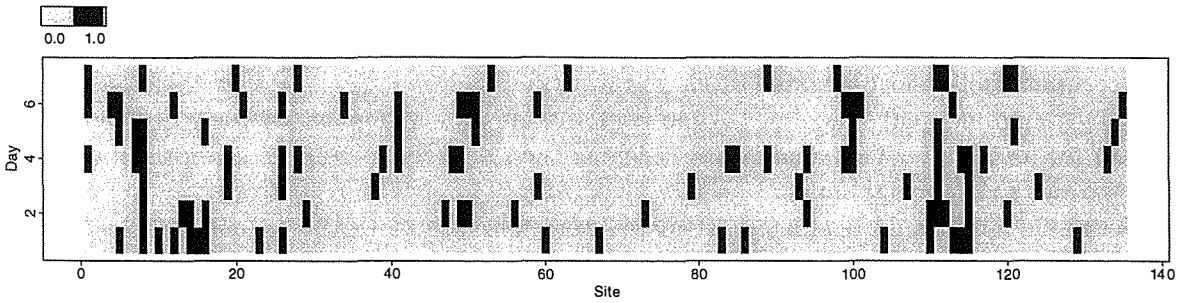


Figure 3.5: The image of dingo presence/absence at the 135 sites over the 7 days; black denotes presence.

We now take into account dingo absence/presence. Consider first modelling the response at a given site with binary responses y_1 and y_2 for locations 1 and 2 respectively. Let q_r be the probability that the chemical at location r attracts a dingo given dingo presence at the site for $r = 1, 2$. Suppose also that given dingo presence, chemicals attract dingoes independently of one another. Let d denote the indicator variable for whether a dingo is present ($d = 1$) or not ($d = 0$) at a site. So certainly if we observe dingoes to be attracted to a chemical or chemicals at a site then d must be equal 1, however, if the observation is negative, no dingoes are attracted, then d is unknown and could be either 0 or 1. Let $s = 0$ if $y_1 + y_2 = 0$, and $s = 1$ if $y_1 + y_2 > 0$, that is, s indicates whether the dingo was observed to be attracted by one or both of the chemicals at the site. Then, for the data (y_1, y_2), the conditional distribution is given by

$$\begin{aligned} p(y_1, y_2 \mid d = 1) &= q_1^{y_1}(1-q_1)^{1-y_1}\, q_2^{y_2}(1-q_2)^{1-y_2}, \\ p(y_1, y_2 \mid d = 0) &= \begin{cases} 1, & s = 0, \\ 0, & s = 1. \end{cases} \end{aligned} \tag{3.2}$$

We assume that marginally d has the Bernoulli distribution with

$$p(d) = p_D^d(1-p_D)^{1-d}, \quad d = 0, 1 \tag{3.3}$$

for probability p_D of dingo presence that does not depend upon which chemicals are at the site. We do note that, above, a suggestion was made in Section 3.3.2 that there may be some interaction between dingo abundance and attractiveness of chemicals, but here we ignore this complex interaction for which there is little information in the data.

The model for (y_1, y_2) given $d = 1$ in (3.2) could be replaced by one that allows for some dependence between chemicals. This essentially amounts to modelling the probabilities of the 2×2 table for $(y_1, y_2 \mid d = 1)$ by using the logit binary model with main effects of chemicals and their interaction, introducing in total a further 15 parameters for interactions, additional to the 6 main effect parameters for chemicals $A, B, \ldots, F$. We have not done this but nevertheless the analysis carried out here remains valid for main effects even if interactions were present.

Conditional Analysis

An initial analysis which avoids estimating dingo presence considers only those site-by-time combinations where dingo presence is known for certain, that is, condition on those sites with s = 1. Then, at a site,

$$p(y_1, y_2 \mid s = 1) = \frac{q_1^{y_1}(1-q_1)^{1-y_1}\, q_2^{y_2}(1-q_2)^{1-y_2}}{1-(1-q_1)(1-q_2)} \tag{3.4}$$

for $(y_1, y_2) = (0,1),(1,0),(1,1)$, which is independent of p_D.

Maximum likelihood estimates of chemical attractiveness $q_A, \ldots, q_F$ can then be obtained by assuming that the site-by-time combinations are independent conditioning on the sites where $s = 1$. The conditional independence is likely to be true except for near temporal and spatial neighbours.

Maximum conditional likelihood estimates were found numerically by using a Nelder-Mead Simplex algorithm with respect to the qs (the only unknown parameters in this case). Approximate standard errors, found by numerically approximating the information matrix, are given in parentheses.

$$\begin{array}{lllllll} \widehat{q}_A & = & 0.478(0.089), & \widehat{q}_B & = & 0.0851(0.041), \\ \widehat{q}_C & = & 0.184(0.070), & \widehat{q}_D & = & 0.538(0.109), \\ \widehat{q}_E & = & 0.406(0.095), & \widehat{q}_F & = & 0.388(0.093). \end{array}$$

As anticipated these are substantially different from the proportions given in equation (3.1) with A best in (3.1) and D best above and estimates larger by a factor of between 4 and 7.

The conditional maximum likelihood estimates are consistent for the qs even if there is dependence between sites and time. A crucial assumption is that there is no interaction between chemicals, *i.e.* the main effects model holds, so that the conditional distribution (3.4) is true at each site and time. The conditional likelihood estimates equate frequencies of the three outcomes (0, 1), (1, 0), (1, 1) with their assumed expected values and are consequently consistent. This follows using the theory of estimating equations; see, for example, Liang and Zeger (1986). Standard errors would be consistent if the assumption of conditional independence were true.

A further reduced analysis might only consider those outcomes where one chemical was effective and the other was not, that is only consider those sites where a distinct preference was shown. Paired comparison techniques, such as a Bradley-Terry statistic, could be used with such a reduced data set. For this dataset, these asymmetric outcomes were observed at only 84 site-by-time combinations, with daily subtotals ranging between 8 and 18. This sample size contains little information on the 15 distinct pairs of chemicals.

Full data model involving dingo presence

We now introduce a new model for the data which is both unconditional, in that it uses all the data, and attempts to accommodate zeros which might be missing data. We construct a model using the possibly unknown dingo presence/absence indicator d for each site. We construct a (full) likelihood and use the EM algorithm (Dempster *et al* , 1977) to obtain estimates of chemical main effects q. We take the so-called complete data to consist of (y_1, y_2, d) for each site-by-time combination. We also assume that the dingo presence variates d are independent over sites and time and drawn from a Bernoulli distribution with p_D constant over regions denoted by sets B_j, in the transect-by-time array, allowing for some systematic change in dingo presence as suggested by Figure 3.4. We return to the choice of the Bs below.

The M stage of the algorithm involves finding maximum likelihood estimates of the chemical effects given all the data (y_1, y_2, d) for all site-by-time combinations. Site-by-time cases contribute independent data so that we consider the complete data log likelihood ℓ

for a site-by-time combination in terms of the q parameters:

$$\ell = \begin{cases} c + d\log(1-q_1) + d\log(1-q_2), & s=0; \\ c + y_1 \log q_1 + (1-y_1)\log(1-q_1) & \\ \quad + y_2 \log q_2 + (1-y_2)\log(1-q_2), & s=1 \end{cases} \tag{3.5}$$

where c is a constant.

For the E-stage of the algorithm we will need to replace d by its expectation given the data. If we denote by $q_A, \ldots, q_F$ the six main chemical effects then the complete data MLE of q_A satisfies

$$\frac{\partial \ell}{\partial q_A} = \sum_i{}^{(A)} \frac{\partial \ell_i}{\partial q_A} = 0$$

where the summation is over all site-by-time combinations where the chemical A is one of the two chemicals at the site. If we denote by y_{A_i} the ith site-by-time response for this chemical and let s_i be the corresponding indicator for no response or some response, then the root to the above equation gives the MLE estimate

$$\widehat{q}_A = \left(\sum_{i:s_i=1}{}^{(A)} y_{A_i} \right) / \left(n^{(A)} + \sum_{i:s_i=0}{}^{(A)} d_i \right) \tag{3.6}$$

where $n^{(A)}$ is the number of site-by-time cases where A is the chemical and where $s_i = 1$. Although at this stage of the analysis d_i is binary, at the E-stage d_i is replaced by a value in $(0,1)$. Then we can think of $\widehat{q}_A$ in (3.6) as being the MLE of q_A from independent binomial data (n', p', y') as follows:

$$\begin{aligned} &\text{for } s_i = 1, \quad \text{data} \sim \text{binomial}(n'=1, p'=q_A, y'=y_{A_i}); \\ &\text{for } s_i = 0, \quad \text{data} \sim \text{binomial}(n'=d_i, p'=q_A, y'=0). \end{aligned}$$

Thus the estimate $\widehat{q}_A$, expression (3.6), is derived by assuming $\sum_{i:s_i=1}{}^{(A)} y_{A_i}$ is binomially distributed with number of trials

$$n^{(A)} + \sum_{i:s_i=0}{}^{(A)} d_i.$$

The complete data also involves d. We assume that the transect-by-time array has been partitioned into disjoint sets $B_j, j = 1, \ldots, K$, where p_{D_j} is then the probability of dingo presence in set B_j.

In our later analysis in Section 3.3.4 we take the sets B_j to be defined as a fixed number of adjacent sites within a day but they could be defined by considering the data and dividing the array into apparently homogenous sets. We have not explored this possibility which, if done automatically, would be akin to edge detection in image analysis and would appear to introduce a higher level of complexity which would not be estimable with such sparse data.

The complete data log likelihood for d is given by the product of terms

$$p(d_i, i \in B_j; p_{D_j}) = \prod_{i \in B_j} p_{D_j}^{d_i} (1-p_{D_j})^{(1-d_i)}$$

for $j = 1, \ldots, K$ assuming independent ds. Thus the log likelihood for d_i for $i \in B_j$ is given by

$$\sum_{i:i\in B_j;\, s_i=1} \log p_{D_j} + \sum_{i:i\in B_j;\, s_i=0} \left\{ d_i \log p_{D_j} + (1-d_i)\log(1-p_{D_j}) \right\} \tag{3.7}$$

and the complete data maximum likelihood estimate of p_{D_j} is given by

$$\widehat{p}_{D_j} = \frac{M^{(B_j)} - M_0^{(B_j)} + \sum_{i:i\in B_j;\, s_i=0} d_i}{M^{(B_j)}} \tag{3.8}$$

where $M^{(B_j)}$ is the number of cases in B_j and $M_0^{(B_j)}$ is the number of cases in B_j with $s_i = 0$.

The E-stage of the algorithm replaces the unobserved d_i in log likelihoods (3.5) and (3.7) by its conditional expectation given current values of parameters and the observed data y . This conditional binary distribution is given by

$$\begin{aligned} p(d = 1 \mid s = 1) &= 1, \\ p(d = 1 \mid s = 0) &= \frac{(1-q_1)(1-q_2)p_D}{(1-p_D) + (1-q_1)(1-q_2)p_D}, \end{aligned} \tag{3.9}$$

so that the required expectation is

$$\bar{d}_i = \frac{(1-q_{1_i})(1-q_{2_i})p_{D_j}}{(1-p_{D_j}) + (1-q_{1_i})(1-q_{2_i})p_{D_j}} \tag{3.10}$$

for $i \in B_j$ and chemicals q_{1_i} and q_{2_i} for site-by-time combination i.
The EM algorithm then proceeds as follows:

1. Obtain initial values for $q_A, \ldots, q_F$ and $p_{D_j}, j = 1, \ldots, K$, and call these the current values.
2. Use current values of qs and p_Ds to obtain $\bar{d}_i$ from equation (3.10).
3. Obtain new values of qs from equations (3.6) and p_Ds from equation (3.8).
4. Check for convergence and, if not, take new values of parameters as current and return to step 2.

The EM algorithm above does allow the estimation to be carried out in a standard statistical package with a generalized linear model facility where the treatment structure can be relatively easily incorporated. Alternatively, direct maximization of the likelihood may be more efficient computationally.

In the next section we consider results for the observed data set.

3.3.4 Results of the basic modeL

Estimates and standard errors

The model was fitted by taking the sets $B_j, j = 1, \ldots, K$ to be defined as adjacent sites within a day. We fitted the model using different sized sets B_j. The EM algorithm behaved well converging in about forty to fifty iterations (results were little changed after 100 and 1000 iterations) for the differently sized B_js. We also calculated the information matrix for the likelihood using the results in Appendix B, and found, surprisingly, that the information matrix was positive semi-definite for only one case, namely B_j constant over nine sites, giving fifteen p_D parameters per day. The estimates and standard errors, calculated from the

asymptotic approximation using the information matrix (having eliminated non-estimable p_Ds), are given below

$$\begin{aligned} \widehat{q}_A &= 0.555(0.164), & \widehat{q}_B &= 0.076(0.044), \\ \widehat{q}_C &= 0.151(0.089), & \widehat{q}_D &= 0.497(0.215), \\ \widehat{q}_E &= 0.413(0.179), & \widehat{q}_F &= 0.323(0.161). \end{aligned}$$

For later reference, parameter estimates are given for the different values of N_s, the number of sites in the sets B_j, in Table 3.7. The estimates above correspond to $N_s = 9$, and estimates appear to be fairly insensitive to different values of N_s, but, as noted above, the information matrix is not consistently positive semi- definite for different values of N_s.

Discussion

Given that for this model potentially one hundred and five p_D parameters are being estimated (actually not all are estimable as for some B_js, all responses are zero), it might be reasonable to assume that the asymptotic standard errors might give poor inferences for the $\widehat{q}$s. This view can be partly confirmed by comparing these estimates and standard errors with those given earlier in Section 3.3.3 for the conditional analysis. The estimates of the qs are about the same size, differing by at most one half standard error, using the standard errors of Section 3.3.3. However, the asymptotic standard errors given above are greater by a factor of two for $\widehat{q}_A$, $\widehat{q}_D$, $\widehat{q}_E$, and $\widehat{q}_F$.

If we compare the conditional analysis of section 3.3.3, which uses only part of the data, with the full data analysis of Section 3.3.3, then the former's assumption of conditional indepencence over sites with s = 1(*i.e.* known to be visited) is supposedly weaker than the unconditional, full data models assumption of independence over all sites.

The analysis of Section 3.3.3 does take account of spatial dependence but the model assumes that the responses at a given site are independent over days. It might be supposed if dingoes had visited a location then they might be more likely to visit it on subsequent days than otherwise, or, conversely, less likely. So, *a priori*, we might suspect some time dependence at a site or location. An alternative way of finding standard errors is the use of the bootstrap and we do this in the next section.

3.3.5 Bootstrap estimates incorporating time dependence

The analysis presented in Section 3.3.3 does provide consistent estimates for the qs provided that the marginal means specified by the likelihood are correct. Attempts to improve the efficiency of the estimates could be made using the method of estimating equations; Liang & Zeger, (1986). This could take into account temporal and/or spatial association. This might also lead to more plausible standard errors but estimates would be solutions of equations with a relatively complex covariance matrix. We refer again to this in Section 3.3.6. Instead we investigate the estimation technique of Section 3.3.3 using the bootstrap when data are resampled from a population which allows some time dependence. Following this approach estimates might not be fully efficient, as the covariance structure is not fully accounted for, but the standard errors should give good inferences.

Hall (1985) suggests ways of resampling a dependent spatial process which incorporates blocking and Hall et al (1995) gives further asymptotic results for time series data. We use these ideas here.

Table 3.8:(a) Empirical probability transition matrix for chemical pair (A, B)

		y_t			
		(0, 0)	(0, 1)	(1, 0)	(1, 1)
y_{t-1}	(0,0)	0.867	0.0	0.111	0.222
	(0,1)	0.0	0.0	0.0	0.0
	(1,0)	0.875	0.0	0.125	0.0
	(1,1)	1.0	0.0	0.0	0.0

(b) Empirical marginal probabilities for y_1

(0, 0)	(0, 1)	(1, 0)	(1, 1)
0.857	0.0	0.127	0.016

We assume that for each chemical pair at a site there is time dependence which is given by a time constant first order Markov chain. Writing y_t for the pair of observations at a site on day t (and dropping the dependence on site), we consider the conditional distribution $(y_t \mid y_{t-1})$ and the subsequent sample of values of $(y_t \mid y_{t-1})$ for each chemical treatment pair for $t = 2, \ldots, 7$. The initial marginal bootstrap distribution for y_1 is obtained from the marginal population of pairs of binary responses for all sites on the first day for the given chemical pair. Alternatively, we could condition on the first days observations all along the transect, maintaining the spatial dependence on the first day.

Thus the first sampling scheme is as follows.

1. For all chemical pairs, (A, B), (A, C) *etc*, construct the conditional sample distribution of $(y_t \mid y_{t-1})$ from the data. Similarly construct the marginal sample distribution of y_1 from the data for all chemical pairs. Both distributions ignore sites.

2. Generate a new set of data by following the original arrangement of chemicals along the transect and drawing each y_1 for each site along the transect from the appropriate chemical pair distribution. Subsequently for each site generate $y_t, t = 2, \ldots, 7$ by drawing y_t from the sample distribution of $(y_t \mid y_{t-1}), t = 2, \ldots, 7$.

In Table 3.8 we give these bootstrap resampling distributions for the chemical pair (A, B) for the conditional distribution $(y_t \mid y_{t-1})$ and the marginal distribution of y_1.

This first sampling scheme does not maintain spatial association in the original data but it does maintain time dependence within a site. Time dependence within a site is obviously confounded with the fixed chemical pair treatment over time, whereas spatial association along the transect is controlled by the balanced allocation of treatments along the transect. Thus, of the two souces of association, the time dependence within sites appears more important to investigate for assessing the precision of the estimates.

The EM method of section 3.3.3 for estimating chemical effectiveness was implemented with bootstrap samples of the data. We found that the EM algorithm required typically about 40 iterations and always less than 200 iterations to obtain convergence to three decimal places but positive semi-definiteness of the information matrix was not checked.

In Table 3.9 we give the resulting bootstrap estimates and standard deviations based on 1000 samples. Bootstrap estimates have standard errors in the range 0.0010 to 0.0026. There is good agreement between estimates in Tables 3.7 and 3.9 with differences of the size

Table 3.9: Bootstrap estimates and standard deviations for the EM analysis using an estimated marginal distribution for the first day (Bootstrap standard error of estimates in range 0.0010 to 0.0026)

Estimate						
N_s	q_A	q_B	q_C	q_D	q_E	q_F
5	0.521	0.075	0.147	0.487	0.389	0.303
9	0.510	0.074	0.148	0.472	0.381	0.301
15	0.507	0.074	0.150	0.475	0.383	0.301
Standard deviation						
5	0.079	0.033	0.051	0.094	0.082	0.072
9	0.082	0.033	0.053	0.097	0.085	0.074
15	0.083	0.033	0.054	0.098	0.086	0.075

Table 3.10: Bootstrap estimates and standard deviation for the EM analysis using observed first day's data

Estimate						
N_s	q_A	q_B	q_C	q_D	q_E	q_F
5	0.533	0.077	0.154	0.492	0.402	0.313
9	0.532	0.075	0.152	0.484	0.391	0.313
15	0.514	0.074	0.151	0.470	0.385	0.306
Standard deviation						
5	0.068	0.031	0.052	0.083	0.072	0.066
9	0.068	0.030	0.051	0.081	0.070	0.064
15	0.071	0.030	0.053	0.089	0.074	0.067

of one or two digits in the second decimal place. The standard deviations in Table 3.9 can be used to estimate the standard errors of the original estimates and these appear relatively stable as N_s varies over 5, 9 and 15.

We also considered the second sampling scheme where the results for the first day y_1 are taken to be the observed data. Results are given in Table 3.10. The estimates are typically larger and standard deviations smaller (as to be expected from theoretical considerations of the bootstrap distributions for y_1) than those of Table 3.9.

Comparing the bootstrap standard errors of Tables 3.9 and 3.10 with the conditional analysis standard errors of section 3.3.3, we note that the former are up to 25% smaller than the latter. Thus the analysis which introduces the dingo presence/absence explicitly does appear to be worthwhile in that estimates of effects are more precise by a factor in the range 20–25% compared with the conditional analysis.

A small theoretical investigation into this question is considered in Appendix A which shows that under simple conditions the conditional analysis loses no efficiency but for a more complex design the conditional analysis is less efficient.

Comparing the bootstrap standard errors of Tables 3.9 and 3.10 with the asymptotic standard errors given in Section 3.3.4, we note that the latter are considerably larger by a

factor of between 1.5 and 2.5, approximately. This suggests that the asymptotic standard errors appear to err on the conservative side quite substantially and are inaccurate.

3.3.6 Further remarks

Modelling spatial smoothness

The basic algorithm assumes that the probability of dingo presence, p_D, remains constant over non-overlapping sets of adjacent sites within each day and changes abruptly at the boundaries of the sets. It would be more appropriate to assume that p_D varied smoothly over the length of the transect and over days at the same site. Additionally, the method of section 3.3.3 might involve a large value of the ratio of the number of parameters to the number of data points. It is well known that the resulting maximum likelihood estimates might be biased in such circumstances. A related situation is explored by Breslow (1981, Section 3.3.2) for investigating the odds ratio in the case where the saturated model is used and the event s = 1, which is $y_1 + y_2 > 0$, is conditioned on; see Section 3.3.3. However, for our analysis, it is suggested otherwise as estimates in Tables 3.7, 3.9 and 3.10 are relatively insensitive to the number of p_Ds being estimated.

Let us introduce explicit notation to indicate the two-dimensional nature of the problem, letting $\{l,t\}$ denote the case at site l along the transect and on day t, which we previously referred to as site-by-time combination i. If we take the sets $B_1,\ldots,B_K$ over which p_D is taken to be constant, to be individual site-by-time cases $\{l,t\}$ then equation (3.8) becomes

$$\widehat{p}_{D_{l,t}} = \begin{cases} 1 & \text{if } s_{l,t} = 1, \\ q_{l,t} & \text{if } s_{l,t} = 0. \end{cases}$$

Silverman *et al* (1990) introduced an adaptation of the EM algorithm which smoothed Poisson means in the context of stereology while Becker and Marscher (1993) used the idea in the context of estimating HIV infection rates from AIDS data. An appropriate adaptation for this example is as follows. For given $\widehat{p}_{l,t}$, let $p^*_{l,t}$ be obtained by smoothing over sites $l-k$ to $l+k$ using weights $w_i, i = 0, \pm 1, \ldots, \pm k$ and $\sum w_i = 1$. Binomial weights

$$w_{i-k} = 2^{-(2k)} \binom{2k}{i}, \quad i = 0,\ldots,2k$$

have been suggested (Silverman *et al* , 1990) and k is chosen by experience with the problem. We can then smooth over time as well to obtain $p^s_{l,t}$ finally as

$$p^s_{l,t} = \alpha p^*_{l,t} + \left(\frac{1-\alpha}{2}\right)\left(p^*_{l,t-1} + p^*_{l,t+1}\right)$$

with α suitably chosen in $[0,1]$ to reflect the time dependence of dingo presence. With only 7 days' data it does not seem appropriate to smooth with a bandwidth greater than 3 days. For smoothing at the boundary either weights can be truncated or data values reflected in the boundary. Here the $p_{l,t}$ are nuisance parameters and not of primary interest and we are only concerned about the effect of the various decisions of smoothing on estimates and standard errors of the q. The resulting $p^s_{l,t}$ are then used in (3.10) to obtain $\bar{d}_{l,t}$.

From extensive numerical studies we have found this approach to be very slow to converge with the number of iterations exceeding 30,000. We do not therefore recommend its use in this context although Anderssen et al (1993) presents compelling theoretical results

and Nychka (1990) suggests that the EM smoothed algorithm is related to a penalised likelihood method. We prefer to model the spatial aspect more directly using Bayesian techniques.

Before considering Bayesian analyses we note Albert and McShane (1995) where generalized estimating equations with spatially correlated binary data are used. This approach could be utilised here by considering the marginal distribution for the data and proposing a suitable correlation structure. Without simplification this would require inversion of a large covariance matrix which could lead to computational difficulties.

Analyses using Bayesian Techniques

An alternative analysis of these data has been carried out using a Bayesian approach and is the subject of the second author's PhD thesis at QUT. Instead of modelling dingo presence/absence at sites to be an independent Bernoulli process with the probability either constant over a small distance (7km) or smoothed, we would prefer to model dingo presence/absence as a dependent binary process. An Autologistic model for dingo presence/absence $d_{l,t}$ at site l and time t is given by

$$p(d_{l,t} \mid \text{all other } d\text{s}) = c \exp\{d_{l,t}[\alpha + \beta(d_{l-1,t} + d_{l+1,t}) + \gamma(d_{l,t-1} + d_{l,t+1})]\}$$

with α related to the marginal probability of dingo presence and β representing dependence on site-neighbours and γ representing dependence on time-neighbours. This model has been used instead of the much simpler independence assumption used here. We report results elsewhere.

Similar problems of having missing data for a binary lattice or map occur with presence/absence maps for animal distribution which are inferred from observed evidence of presence. Hogmander and Moller (1995) provide details of Bayesian methods not dissimilar to our proposed Bayesian analysis for the dingo data using a single parameter Ising model.

Conclusion

In conclusion, we have considered various analyses for the data. A conditional analysis provided estimates and standard errors, which, in the context of the other analyses, are plausible. A full data model, introducing the notion of dingo presence, provided plausible estimates but standard errors from asymptotic approximations were too large. A bootstrap resampling scheme taking into account probable time dependence gave standard errors smaller than other approaches and appeared worthwhile. Additionally, the reduction in size of standard errors, giving greater precision of estimates, made worthwhile the development of the more complex model.

We note that the absence/presence full data model can be straight forwardly extended or modified to involve more than bivariate responses and to Poisson or other discrete non-negative distributions.

Acknowledgements

The authors acknowledge the useful remarks of referees on earlier versions of this paper which led to substantial improvements. We gratefully acknowledge the assistance of Ms Michele Haynes with some of the computing work. The data were collected by staff of the Queensland Lands Department and Dr Chris King of QUT to whom we are indebted. The data can be obtained from the web site *http://www.math.fsc.qut.edu.au/papers/*.

References

Albert, P.S. and McShane, L.M. (1995). *Generalized estimating equations approach for spatially correlated data: applications to the analysis of neuroimaging data.* Biometrics 51, 627–638.

Anderssen, R.S., Latham, G. and Westcott, M. (1993). *Statistical methodology for Inverse Problems*, pp 1–7 in 'Stochastic Models in Engineering, Technology and Management' eds. S. Osaki and D.N. Pra Murthy. Singapore: World Scientific.

Becker, N.G. and Marschner, I.C. (1993). *A method for estimating the age-specific relative risk of HIV infection from AIDS incidence data.* Biometrika 80, 165–78.

Breslow, N. (1981). *Odds ratios when data are sparse.* Biometrika 68, 73–84.

Corbett, L. (1995). *The dingo in Australia and Asia.* Sydney: Uni NSW Press.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). *Maximum likelihood from incomplete data via the EM algorithm.* J.R. Statist. Soc. B 39, 1–38.

Graybill, F.A. (1983). *Matrices with applications in statistics*, 2nd edition. Wadsworth: Belmont, California.

Hall, P. (1985). *Resampling a coverage process.* Stoch.Proc.Applic. 20, 231–46.

Hall, P., Horowitz, J.L. and Jing B-Y (1995). *On blocking rules for the bootstrap with dependent data.* Biometrika 87, 561–74.

Hogmander, H. and Moller, J. (1995). *Estimating Distribution Maps from Atlas Data using Methods of Statistical Image Analysis.* Biometrics 51, 393–404.

Liang , K-Y. and Zeger, S.L. (1986). *Longitudinal data analysis using generalized linear models.* Biometrika 73, 13-22.

Louis, T. A. (1982). *Finding the observed information matrix when using the EM algorithm.* J.R. Statist. Soc. B, 44, 226-233.

Mitchell, J. (1988). *Animal Attractant Project.* Brisbane: Rural Lands Protection Board.

Nychka, D. (1990). *Some properties of adding a smoothing step to the EM algorithm.* Statist. Probab. Lett. 9, 187–193.

Silverman, B.W., Jones, M.C., Wilson, J.D. and Nychka, D.W. (1990). *A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography.* J.R. Statist.Soc.B 52, 271–324.

Welsh, A.H., Cunningham, R.B., Donnelly, C.F. and Lindenmayer D.B. (1996). *Modelling the abundance of a rare species; statistical models for counts with extra zeros.* Ecological Modelling 88, 297-308.

3.3.7 Appendix A

Information for unconditional/conditional analysis

Here we investigate the information lost due to the conditional analysis. Assuming independence over sites and times the likelihood can be written using

$$p(y) = p(y,s) = p(y \mid s)p(s)$$

as follows

$$p(y;q,p_D) = \prod_{i:s_i=s;j=0,1} p(y_i \mid s_i = s;q,p_D)\, p(s_i = s;q,p_D) \tag{3.11}$$

Now

$$p(y \mid s=0;q,p_D) = \begin{cases} 1, & \text{if } (y_1,y_2) = (0,0) \\ 0, & \text{otherwise.} \end{cases}$$

Also

$$p(y \mid s=1;q,p_D) = p(y \mid s=1;q)$$

and does not involve p_D. Additionally

$$p(s;q,p_D) = (ap_D)^s(1-ap_D)^{1-s}, \quad s=0,1$$

with

$$a = 1-(1-q_1)(1-q_2).$$

So the likelihood $p(y;q,p_D)$ simplifies to give

$$p(y;q,p_D) = \prod_{i:s_i=1} p(y_i \mid s_i=1;q) \times \prod_i (a_ip_D)^{s_i}(1-a_ip_D)^{1-s_i} \tag{3.12}$$

Consider first a simpler experiment which trials only two chemicals and these two chemicals are present at each site. We assume p_D is constant and known over all site-by-day combinations of which there are N. On average we expect only Nap_D site-by-time combinations to have $s=1$ and below we determine the information which is lost by ignoring the information from the second term in (3.12).

For this case, $a = 1-(1-q_1)(1-q_2)$ for all cases and the second term in (3.12) is a product of Bernoulli terms with statistics s_i and probability of success ap_D. Thus the information matrix for q_1, q_2 and p_D derived from $p(s)$ is of rank 1. We find for N combinations of site-by-time and parameters q_1, q_2 and p_D that

$$\mathrm{E}\left[\text{information for } p(y|s=1)\right] = \frac{Np_D}{a}\begin{pmatrix} -\frac{q_2}{q_1(1-q_1)} & -1 & 0 \\ -1 & \frac{q_1}{q_2(1-q_2)} & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

and

$$\begin{aligned} &\mathrm{E}\left[\text{information for } p(y)\right] \\ &= \mathrm{E}\left[\text{information for } p(y|s=1)\right] \\ &+\frac{Np_D}{a}\frac{1}{1-ap_D}\begin{pmatrix} (1-q_2)^2 & (1-q_1)(1-q_2) & K \\ (1-q_1)(1-q_2) & (1-q_1)^2 & K \\ K & K & K^2 \end{pmatrix} \end{aligned}$$

for some positive constant K. It can be easily shown, *e.g.* Graybill (1983, p184), that in this case the determinant of the submatrix corresponding to q_1 and q_2 of the inverse of the information matrix for $p(y)$ is equal to the inverse of the information matrix of q_1 and q_2 for $p(y \mid s = 1)$. This indicates that the conditional analysis based on $s = 1$ loses no information for q_1 and q_2 compared with the unconditional analysis although $p(s)$ contains information about q_1 and q_2. Here, essentially the information for q_1 and q_2 in $p(s)$ is confounded with that for p_D which is being estimated. However, in this simpler experiment if sites were allocated two chemicals of the same type as well as different chemicals then this data would not confound q_1 and q_2 with p_D. For, suppose a block consists of three pairs of chemicals at the three sites where the pairs are chemicals $(1,1),(2,2)$ and $(1,2)$. Suppose also s_{ij}, $(i,j) \in \{(1,1),(2,2),(1,2)\}$ denote the s-outcomes for the three sites. The score for q_1

$$\frac{\partial}{\partial q_1} \log p(s)$$

with $s = (s_{11}, s_{22}, s_{12})$, involves the statistics s_{11} and s_{12}, that for q_2, s_{22} and s_{12}, and that for p_D, s_{11}, s_{22} and s_{12}. These three scores are linearly independent and hence the expected information matrix derived from $p(s)$ for q_1, q_2 and p_D is of full rank, namely three in this case. Thus the overall information for all parameters will be increased by utilising $p(s)$.

In this second case therefore, unlike the first, the unconditional analysis has greater information than the conditional. For the actual experiment carried out, there are six chemicals used in fifteen different distinct pairs. There are therefore fifteen s-statistics from which the six plus one or seven score functions are derived to provide information for $q_A, \ldots, q_F$ and p_D from $p(s)$. Here the information matrix for $p(s)$ is full rank and the unconditional analysis provides additional information over that just based on sites where dingo presence is certain.

3.3.8 Appendix B

Standard Errors for EM Algorithm Estimates

Louis (1982) shows that the observed information matrix for the observed data can be obtained from the complete data as follows:

$$\frac{-\partial^2 \ell_{\text{obs}}}{\partial\theta^{\mathsf{T}}\partial\theta} = \mathrm{E}\left[\frac{-\partial^2 \ell_{\text{com}}}{\partial\theta^{\mathsf{T}}\partial\theta}\right] \mathrm{Var}\left[\frac{\partial \ell_{\text{com}}}{\partial\theta}\right]$$

where ℓ_{obs} and ℓ_{com} are the observed data and complete data loglikelihoods respectively, and expectation is with respect to the missing data conditional on the observed. Here θ consists of the q and p_D parameters and the ds are the missing data.

We first find the terms $\frac{\partial}{\partial\theta}\ell_{\text{com}}$, combining (3.2) and (3.3) to obtain $\exp(\ell_{\text{com}})$ for a case. Then, for example,

$$\frac{\partial \ell_{\text{com}}}{\partial q_A} = \sum_{i:s_i=1}^{(A)} \frac{y_i - q_A}{q_A(1-q_A)} + \sum_{i:s_i=0}^{(A)} \frac{-d_i}{1-q_A}$$

so

$$\mathrm{Var}\left[\frac{\partial \ell_{\text{com}}}{\partial q_A}\right] = \sum_{i:s_i=0}^{(A)} \frac{\mathrm{Var}[d_i]}{(1-q_A)^2}$$

and

$$\mathrm{Cov}\left[\frac{\partial \ell_{\text{com}}}{\partial q_A}, \frac{\partial \ell_{\text{com}}}{\partial q_B}\right] = \sum_{i:s_i=0}^{(A,B)} \frac{\mathrm{Var}[d_i]}{(1-q_A)(1-q_B)}$$

where the summation is over those cases where both chemicals A and B are located. We note

$$\mathrm{Var}[d_i] = \frac{p_{D_j}(1-p_{D_j})(1-q_A)(1-q_B)}{\left\{(1-p_{D_j}) + (1-q_A)(1-q_B)p_{D_j}\right\}^2} \tag{3.13}$$

$$= \bar{v}_i, \text{say}, \tag{3.14}$$

for chemicals A and B located at case i and the appropriate p_{D_j}. Now

$$\frac{\partial \ell_{\text{com}}}{\partial p_{D_j}} = \sum_{i:i\in B_j;\, s_i=0} \frac{d_i - p_{D_j}}{p_{D_j}(1-p_{D_j})} + \sum_{i:i\in B_j;\, s_i=1} \frac{1}{p_{D_j}}$$

so

$$\mathrm{Var}\left[\frac{\partial \ell_{\text{com}}}{\partial p_{D_j}}\right] = \sum_{i:i\in B_j;\, s_i=0} \frac{\bar{v}_i}{p_{D_j}^2(1-p_{D_j})^2}$$

and

$$\mathrm{Cov}\left[\frac{\partial \ell_{\text{com}}}{\partial p_{D_j}}, \frac{\partial \ell_{\text{com}}}{\partial p_{D_{j'}}}\right] = 0, \quad j \neq j'.$$

Also

$$\mathrm{Cov}\left[\frac{\partial \ell_{\text{com}}}{\partial q_A}, \frac{\partial \ell_{\text{com}}}{\partial p_{D_j}}\right] = -\sum_{i:i\in B_j;\, s_i=0}^{(A)} \frac{\bar{v}_i}{(1-q_A)p_{D_j}(1-p_{D_j})}.$$

The matrix

$$\mathrm{Var}\left[\frac{\partial \ell_{\text{com}}}{\partial\theta}\right]$$

is given by the above terms. We note

$$\mathrm{E}\left[\frac{-\partial \ell^2_{\mathrm{com}}}{\partial q_A^2}\right] = \sum_{i:s_i=0}^{(A)} \frac{\bar{d}_i}{(1-q_A)^2} + \sum_{i:s_i=1}^{(A)} \left\{\frac{y_{li}}{q_A^2} + \frac{(1-y_{li})}{(1-q_A)^2}\right\}$$

$$\mathrm{E}\left[\frac{-\partial \ell^2_{\mathrm{com}}}{\partial {p_{D_j}}^2}\right] = \sum_{i:i\in B_j;\ s_i=1} \frac{1}{p_{D_j}^2} + \sum_{i:i\in B_j;\ s_i=0} \left\{\frac{\bar{d}_i}{p_{D_j}^2} + \frac{(1-\bar{d}_i)}{(1-p_{D_j})^2}\right\}$$

and all other terms are zero so that the matrix

$$\mathrm{E}\left[\frac{\partial^2 \ell_{\mathrm{com}}}{\partial \theta^{\mathsf{T}} \partial \theta}\right]$$

is diagonal.

3.4 Discussion

Although some useful results have been obtained using this approach, it has also suffered a number of difficulties, namely, in the way that spatial dependence was modelled, in estimation of standard errors, and in interpretation of results.

A major drawback has been that the underlying spatio-temporal process had to be treated in an *ad hoc* fashion due to the constraints (conceptual and computational) imposed by the frequentist approach, rather than as an integral part of the model. Spatial dependence was modelled in one of two ways: by estimating probabilities of presence in blocks of site-time; and by smoothing estimates of probabilities of presence along sites. The first blocked approach suffers from the typical problem of how to select block-size. It also does not reflect the continuous nature of the changes in dingo presence along the transect and over time. The second approach allows very limited smoothing of presence probabilities over sites, within a given time period. Neither of these approaches allowed modelling of the spatial and temporal patterns in dingo presence. In addition, spatial and temporal dependence of dingo presence would also be modelled contemporaneously. Suitable models for binary spatio-temporal patterns will be the subject of the next chapter (Chapter 4), which provides the basis for the Bayesian approach explored in the rest of the thesis.

Since the asymptotic estimates of standard errors appeared to be too large, bootstrapped estimates of standard errors for both the blocked and smoothed estimates of dingo presence were investigated. A full hierarchical model would allow estimation of standard errors at the modelling stage rather than as an *ad hoc* step. Some methods of constructing such a hierarchical model use a Bayesian approach and are the topic of investigation in the rest of the thesis.

Furthermore, although point estimates and standard errors were obtained using maximum likelihood, these are only two-number summaries based on the likelihood of the test statistics given the data. A Bayesian approach is able to supply a full posterior distribution of model parameters and monitoring statistics.

Chapter 4

Binary Markov Random Fields

Contents

4.1 Introduction 65
4.2 Markov Random Field Models 66
4.2.1 Neighbourhood 67
Neighbourhood order 67
Neighbour labels 67
4.2.2 Cliques 68
4.2.3 Asymptotics 69
4.2.4 Edge Sites 70
4.2.5 Markov property 73
4.2.6 Equivalence between Gibbs and Markov random fields 74
4.2.7 Specific Markov random field models 76
Verhagen's model 76
Pairwise Interaction General model 76
Auto-binomial model 77
Autologistic model 77
Derin-Elliot model 78
Auto-Poisson model 78
Auto-Gaussian model 78
4.3 Anisotropic Autologistic model 78
4.3.1 Definition of Ising Model 80
4.3.2 Definition of Autologistic Model 81
4.3.3 Isotropy 82
4.3.4 Phase Transition 84
4.3.5 Theoretical results for the Ising model 85
Dual representation 87
Bragg-Williams approximation 88
Bethe-Peierls approximation 90
Onsager's method and other exact methods 92
Relationship to the Gaussian model 93
Hyperscaling hypothesis 94

4.4 Simulation from MRFs 95
4.4.1 Definition of Markov chain Monte Carlo 95
4.4.2 Metropolis-Hastings 96
Proposals for $[0,1]$ data 97
Rotated uniform distribution 97
Truncated uniform distribution 98
Normal distribution 99
Truncated normal distribution 100
4.4.3 Gibbs Sampling 100
4.4.4 Hybrid strategies 101
4.4.5 Other samplers 102
4.4.6 MCMC performance: standard error 104
IACT: Naive estimator 105
IACT: Truncated periodogram estimator 106
IACT: Spectral density estimator using Bartlett window 106
Run length 107
4.4.7 Diagnostics 108
Discussion 111
4.5 Statistical inference for MRFs 111
4.5.1 Reparameterization 112
4.5.2 Maximum Likelihood Estimation 113
Asymptotic Maximum Likelihood 114
Approximate Maximum Likelihood 115
Maximum Pseudolikelihood 116
4.5.3 Minimum χ^2 Estimation 117
A Least Squares Solution 119
Neighbourhood basis estimation 119
4.6 Discussion 119

4.1 Introduction

> I never could make out what those damned dots meant.
> - *Lord Randolph Churchill: on decimal points, in W. S. Churchill "Lord Randolph Churchill" (1906).*

This chapter begins the investigation of the Bayesian approach to the problem outlined in the motivating Chapter 2: how to model ambiguous presence/absence data by separating the data model for the binary response from an underlying presence/absence process. In the spatial or spatio-temporal context, the underlying process is generally or necessarily, respectively, in more than one dimension. The review of binary spatial data models suitable for describing these underlying processes (Section 2.4) highlighted Markov random fields (MRFs) as potential candidates for describing spatio-temporal dependence within a hierarchical framework. This chapter provides the theoretical basis required for implementation of this approach in Chapters 5, 6, 7.

MRFs derive from random fields (Section 4.2), whose fundamental characteristics are neighbourhoods, cliques, asymptotics and boundary effects. Random fields are defined as a prelude to introducing a well-established global model for the joint distribution of a spatially dependent map, called a Gibbs Random Field (Section 4.2.4). If this conditional distribution only depends on neighbours, then the field is said to have the Markov property (Section 4.2.5). This equivalence between Gibbs random fields and MRFs means that spatial dependence on a lattice can be modelled both jointly or conditionally (Section 4.2.6). This facilitates simulation and inference without compromising the spatial dependence assertion.

Due to difficulties with deriving properties of these distributions analytically, inference is based on simulation methods (Section 4.4). Approaches to inference for MRFs in the statistical literature have various advantages and disadvantages, as outlined in Section 4.5.

In this thesis a particular binary MRF, the autologistic model, is used to illustrate theory and is introduced in Section 4.3. This model is suitable for the *dingo* case study (Chapters 5 and 7), and therefore by extension to other applications discussed in Section 2.3. Some specific applications of the model have previously been outlined in Section 2.4. These cover diverse fields of study such as statistical physics, image analysis, biogeography and economics.

The one-parameter autologistic model is currently popular in the spatial statistics literature (Sections 2.3.2 and 2.3.3) due to its simplicity. Here we consider an extension to this model, the three-parameter autologistic model, which is able to model prevalence as well as spatial interaction. It also separates spatial interaction into two components, such as horizontal *vs* vertical, east-west *vs* north-south, or one-dimensional space (transect) and time. This version is useful for the *dingo* case study and other spatial statistics applications.

The autologistic bears close resemblence to a physical model called the Ising model, presented in Section 4.3.1. Many of the Ising model's properties, and therefore those of the autologistic, have been investigated for different reasons in the statistical physics literature; these are presented in Sections 4.3.4 and 4.3.5. One of the characteristics of the Ising/autologistic model is the existence of a phase transition. For some values of spatial interaction parameters, the distribution of the difference between the number of presences and absences on the lattice is bimodal. Thus, for critical parameter values, configurations with large patches of presence are highly likely, yet configurations with large patches of absence are also highly likely. Hence this model is suitable for modelling systems where, once spatial interaction passes a critical point, lattice sites begin to behave in a cooperative fash-

ion and exhibit long range dependence. Gaussian MRFs do not exhibit this phenomenon and so in these situations, the autologistic should be used.

The autologistic/Ising model has been utilized in the disciplines of statistical physics, image analysis and statistics. This has led to diverse terminology, notation and some concepts. In this thesis I adopt a mixture of both terminologies: statistical physics terminology is used to ensure that results from the literature can be referred to without too much trouble; and statistical terminology is used to interpret statistical physics results and present all phases of statistical modelling and inference. Alternative terminologies are indicated in footnotes unless significant to presentation of results.

4.2 Markov Random Field Models

In this section random fields are defined in conjunction with their fundamental properties of neighbourhood, cliques, asymptotics and boundary effects. Random fields are important to the latter half of the thesis, beginning with Chapter 5. Markov random fields, a subset of the larger group of Random fields, are suitable for modelling the underlying spatio-temporal process. A particular type of random field model, the Autologistic, is used for application to the *dingo* case study in all remaining chapters.

Here we consider a multivariate random variable[1] $z = \{z_i : i \in \mathcal{L}, z_i \in \Omega_0\}$ measured on a 2-dimensional rectangular lattice, where the random variables z_i are indexed by some *lattice index set* $\mathcal{L}$, and take on values which are contained in some *individual sample space* Ω_0. The *overall sample space* for the collection z is $\Omega = \Omega(\mathcal{L}) = {\Omega_0}^{|\mathcal{L}|}$. Individual positions i in the lattice are called sites. For binary random fields z the random variables $\{z_i\}$ each have only two permissible values, *i.e.* $|\Omega_0| = 2$. In the applications of interest, the two values are $\{0, 1\}$ representing absence and presence respectively[2]. When $|\Omega_0| > 2$ then we have categorical[3], count (Ickstadt & Wolpert 1999) or continuous (Weir & Pettitt 1999, Cressie 1993) random fields. These cases are beyond the scope of this thesis, which focuses only on the presence/absence case.

For z observed on a 2-dimensional rectangular lattice, which has n_1 rows and n_2 columns, a natural index set is the coordinate system $\mathcal{L}^* = \{i^* = (i_1, i_2) : i_1 = 1, 2, \ldots, n_1; i_2 = 1, 2, \ldots, n_2\}$. Simplified notation for indices is obtained by using subscripts $\mathcal{L} = \{i : i = 1, 2, \ldots, L\}$, with $L = n_1 \times n_2$. To transform from the first $i^* = (i_1, i_2)$ to the second i index system, I adopt the programming device $i = n_2(i_1 - 1) + i_2$. The form of the index only becomes relevant in programming and computation, so we adopt the less cluttered i version in most instances, such as general derivation of results. On discussion of computational issues the more detailed indexing system is often more useful, for example where i represents spatio-temporal location st.

Markov random fields are members of the exponential family, can be highly multi-dimensional and are able to describe systems with a high degree of dependence. For members of the exponential family, let $z \in \Omega$ denote the random variable observed on lattice $\mathcal{L}$, with components for each site i denoted by z_i indexed by $i \in \mathcal{L}$. Then the probability density of

[1] Multivariate z is variously termed a map (geography), configuration (physics), or a colouring (image analysis).

[2] or when modelling magnets $\Omega_0 = \{-1, +1\}$ represents up and down spins respectively.

[3] Random variables on the lattice (pixels in an image) may take on any of a fixed number of values (colours or grey levels in an image).

z depends on parameter vector θ and can be written in the form

$$\begin{aligned} p(z|\theta) &= h(z,\theta)/c(\theta) \\ h(z,\theta) &= \exp\{\phi(\theta)^{\mathsf{T}} V(z) + \psi(z)\} \end{aligned} \tag{4.1}$$

where $\phi(\theta)$ is known as the canonical transformation of the parameter θ; $V(z)$ is the canonical statistic (transformations of the data); $\psi(z)$ is a function of the data only; and $c(\theta)$ is a function of the parameter only. In these distributions the function $c(\theta)$ depending only on θ can be viewed as the normalization constant of the unnormalized density $h(z,\theta)$:

$$c(\theta) = \sum_{z \in \Omega} h(z,\theta). \tag{4.2}$$

In the sections to follow we refine the general exponential family definition to the specific family of Markov Random field models.

4.2.1 Neighbourhood

The concept of neighbourhood is central to the definition of MRF models, which will be used to describe the underlying spatio-temporal presence/absence process. Small localized neighbourhoods around individual sites form the building blocks that we use to piece together a global picture of spatial pattern over the entire space of interest. It is important to define the size and shape of the neighbourhood $\mathcal{N}$ assumed in a MRF model since this defines the probability density.

Commonly, the neighbourhood size and shape are described using a single integer called the *order of neighbourhood*, as defined in Section 4.2.1.

In some models the position of a neighbour relative to the central site may be important. For example the soil gradient may cause diagonal neighbours to be more spatially related than others, or an image texture may exhibit vertical striations as defined on page 35. These are called anisotropies in the model, and affect the notation of neighbours and parameters. In Section 4.2.1 I outline a simple integer labelling system and a more complex neighbourhood basis system to differentiate neighbour positions, both selected from the image analysis literature. These extend the simple $i \sim j$ notation prevalent in the statistical literature.

Neighbourhood order

Neighbourhood order is a common definition of neighbourhood size and shape found in the literature (Besag 1974, Dubes & Jain 1989). In Figure 4.1 we see a generic neighbourhood centred on a site i located in the centre of the grid. Neighbouring cells are labelled with their *order*.

For example, a neighbourhood of order 1 only includes neighbouring sites directly above (N) and below (S), and to the left (W) and right (E) of a central site; a neighbourhood of order 2 extends to include diagonal neighbouring sites to the NE, NW, SE and SW. The order 3 neighbourhood includes all neighbours up to order 2, in addition to sites within 2 units in any of the four cardinal directions, N, S, E and W.

Neighbour labels

I have rearranged the integer labels used by Dubes & Jain (1989) and Chen (1988) within geometric orders to make the labelling more systematic. In this system I have chosen

(a)

5	4	3	4	5
4	2	1	2	4
3	1	0	1	3
4	2	1	2	4
5	4	3	4	5

Figure 4.1: With respect to the central site indicated by 0, the geometric order of neighbourhood is provided for nearby neighbours.

arbitrarily to give E precedence over W, S over N, and SE and SW over the other diagonal directions, with numbering following geometric order.

Figure 4.2, based on Figure 2-1 in Chen (1988), shows how the notation $i : +\delta$ can be used to index neighbouring sites in directions defined on the grid relative to a central site i. For example, $\delta \in \{+1, -1\}$ corresponds to horizontal neighbours to the right and left (respectively) of site i, and $\delta \in \{+3, -3\}$ denote neighbours on the positive diagonal. Let Δ denote the set of neighbouring site indices, *e.g.* $\delta \in \Delta = \{\pm 1, \pm 2, \pm 3, \pm 4\}$ for a second-order neighbourhood. Also site $i : +1$ is to the East of centre, or 1 unit to the right; and site $i : -9$ is to the NWN of centre, or 2 units up and 1 to the left.

(b)

-11	-9	-6	-10	-12
-8	-3	-2	-4	+7
-5	-1	0	+1	+5
-7	+4	+2	+3	+8
+12	+10	+6	+9	+11

Figure 4.2: With respect to the central site i denoted 0 on these grids, this table shows notation for indexing sites in neighbourhood to site i.

A more flexible way to identify the neighbours of a particular site is to define a neighbourhood basis. This is prompted by Possolo (1986*b*)'s definition of a $\mathcal{N}$-neighbourhood and uses the idea of a basis taken from Linear Algebra. A "reference" neighbourhood is created, and then used as a template to generate neighbourhoods of any site. A *neighbourhood basis* $\mathcal{N} \subset \mathcal{L}$ is centred at but does not include an arbitrary origin site $i_\circ \in \mathcal{L}$ (preferably away from the boundary of $\mathcal{L}$). Then the *$\mathcal{N}$-neighbourhood* of site $i \in \mathcal{L}$ is

$$\mathcal{N}(i) = \{i : +\delta \text{ where } \delta \in \mathcal{N}\}, \tag{4.3}$$

and in general, of a finite group of sites $I \subset \mathcal{L}$ is

$$\mathcal{N}(I) = \bigcup_{i \in I} \mathcal{N}(i) - I. \tag{4.4}$$

Thus $j \in \mathcal{N}(i)$ is a neighbour of site i. The advantages of the neighbourhood basis notation is that it explicitly and transparently refers to both distance and orientation from the centre and extends easily to a non-lattice context.

4.2.2 Cliques

Cliques expand on the definition of neighbourhood in order to define the probability distribution of Markov random field models. Clique equivalence classes underlie a fundamental

result giving the unique correspondence between a global model of a map (Gibbs random field) and its local conditional counterpart (Markov random field), investigated in Section 4.2.6.

Sites can be classified into broad groups called *cliques* (Dubes & Jain 1989, Besag 1974, Possolo 1986*a*) where sites in each group are neighbours of all other sites in the group. For a given neighbourhood basis $\mathcal{N}$, all sites in a *clique* $C \in \mathcal{L}$ are neighbours of all other sites in that clique:

$$i \in C \iff i \in N(j) \qquad \forall j \in C \tag{4.5}$$

We can focus on cliques containing a particular site, and thus generate the set of all cliques from our sampling window. With respect to some neighbourhood basis $\mathcal{N}$, denote by $C(\{i\})$ the set of all cliques which contain site i is

$$C(\{i\}) = \{C : i \in C\} \quad \text{and let} \quad \mathcal{C} = \bigcup_{i \in L} C(\{i\}) \tag{4.6}$$

be the collection of all cliques on the sampling window $L \subset \mathcal{L}$.

Note that by definition, cliques for a site i may be overlapping and so are not disjoint. It is useful to focus on just those cliques whose sites have a particular neighbour relationship (*e.g.* all cliques containing horizontal neighbours.) Define the kth equivalence class of cliques as the set of all cliques whose sites bear the same relationship as that between the origin and the kth element (or group of elements) of the neighbourhood basis $\mathcal{N}$. Let $\mathcal{N}_k \subset \mathcal{N}$ denote a subset of elements in the neighbourhood basis. The kth equivalence class of cliques is defined as:

$$\mathcal{C}_k \equiv \mathcal{C}_k(\mathcal{N}) = \{C \in \mathcal{C} : \forall i, j \in C \quad \exists \delta \in \mathcal{N}_k \quad \text{s.t.} \quad i - j = \delta - i_\circ\}. \tag{4.7}$$

Let $\mathcal{C}_0$ be the class of all single sites in the lattice $\mathcal{L}$, *i.e.* the set of all single-site cliques. A neighbourhood system Θ can be formed (Besag 1974, Possolo 1986*b*, Dubes & Jain 1989) by partitioning the cliques into equivalence classes $\{\mathcal{C}_k\}_{k=0}^{K}$.

4.2.3 Asymptotics

In order to understand the effect of edge sites (next section) on inference, it is essential to first understand asymptotics for Markov random field models.

In the spatial context, asymptotics can be investigated by gradually increasing the spatial coverage of sampling windows (Pickard 1976, Kindermann & Snell 1980, Ogata & Tanemura 1984), rather than the intensity of sampling. Data observed within a finite sampling window L_n are embedded within an infinite sample space Ω. Sandwiching sampling windows L_n between sequences of circles or rectangles in $\Re^2$, and equivalently in higher dimensions, ensures they increase in a regular way towards Ω. Define a *circle* A of radius $0 \leq r < \infty$ centred at index $i_{N_\circ} \in \mathcal{L}$ as the set

$$A = \{i \in \mathcal{L} : d(i, i_{N_\circ}) \leq r\} \tag{4.8}$$

where in this thesis we assume a Euclidean metric d on $\mathcal{L}$.

The following derivation is based on Possolo (1986*b*). Now consider the sequence of sampling windows $\{L_n : n = 1, 2, \ldots\}$ which comprise a sequence of finite subsets of the index set.

$$L_n \subset L_{n+1} \quad \text{where} \quad L_n \subset \mathcal{L}, \tag{4.9}$$

with limit being the entire lattice:

$$\bigcup_{n=1}^{\infty} L_n = \mathcal{L}. \tag{4.10}$$

Then this sequence $\{L_n\}$ is said to *increase regularly* to $\mathcal{L}$ if

1. There exists sequences $\{A_n\}, \{B_n\}$ of circles in $\mathcal{L}$ all centred at the same fixed point $i_{N_0} \in \mathcal{L}$ such that these sphere sequences are monotonic increasing

$$A_n \subset A_{n+1}, \qquad B_n \subset B_{n+1} \tag{4.11}$$

with limit being the entire lattice

$$\bigcup_{n=1}^{\infty} A_n = \mathcal{L}, \qquad \bigcup_{n=1}^{\infty} B_n = \mathcal{L}; \tag{4.12}$$

2. The sampling window is 'sandwiched' between these two circular sequences $A_n \subset L_n \subset B_n$ in a finite manner

$$\sup_n \left| \frac{B_n}{A_n} \right| < \infty. \tag{4.13}$$

We will denote this by $\{L_n\} \xrightarrow{\text{i.r.}} \mathcal{L}$ where i.r. refers to the property of *increasing regularly* as defined above.

The precise reason for phase transition is related to asymptotics. In particular, an infinite MRF (local model) which is restricted to a finite lattice is not the same as a GRF (global model) applied to a sublattice. Firstly, the global GRF model might not exhibit the local stationarity required, and secondly the the local Markovian property of a MRF may fail at the boundaries. However, the GRF and MRF may be equivalent for some parts of the parameter space. Since an infinite MRF is actually a limit of a finite GRF as the lattice increases regularly, differences between the models are only due to boundary effects.

4.2.4 Edge Sites

A decision on how to handle edge sites is essential with any practical investigation, such as those presented in Chapters 5–7.

Edge sites for any set of sites I are its neighbours $\mathcal{N}(I)$. Thus the edge sites of a sampling window L_n are the neighbours around its border $\mathcal{N}(L_n)$. The number of edge sites increases with the size of the sampling window. For a lattice of dimension $\nu \geq 2$, as the sampling window expands the number of edge sites grows to infinity (Pickard 1982):

$$|\mathcal{N}(L_n)| = O(|L_n|^{\nu-1}) \quad \text{as} \quad L_n \xrightarrow{\text{i.r.}} \mathcal{L} \tag{4.14}$$

The rate of expansion is slowest for the two-dimensional case examined in this thesis, where the number of border sites increases linearly with lattice size.

Shapes of finite lattices affect results depending on ratios of linear dimensions (Binder & Heermann 1997). For example on the 2D lattice, this ratio is of the number of rows to number of columns.

Edge sites are essentially *missing* variables. There are two major approaches to dealing with edge sites which parallel the ways in which missing values are dealt with in statistics.

The first approach is to *condition* on the edges, and proceed with analysis as though their values are known. Possolo (1986*b*) advocates, following the general principles of Anderson (1970), that inference on spatial interaction parameter β should be conditional on minimal sufficient statistics for nuisance parameters (or edge sites) $z_{\mathcal{N}(L_n)}$. This is similar to Ripley (1981)'s use of a "guard area" around the sampling window. An ad-hoc analysis of sensitivity to the conditioned edge site values would involve re-running the analysis for different 'representative' conditioning values and summarising their effect on conclusions.

The second approach is to attempt to *estimate* the missing values (in the tradition of Dempster et al. (1977)'s EM algorithm) following any of a variety of methods. These assume one of the following.

- The lattice is a finite island and has no edge sites.
- The lattice is wrapped onto a torus so that neighbours of edge sites are determined by sites on the opposite border.
- The edge sites are missing values of some latent variables which need to be estimated via some technique, for example: reflection at the boundary; nearest neighbour average; bootstrapping the joint distribution; and bootstrapping the conditional distribution.

A popular choice in the literature (*e.g.* Domb & Green (1972*a*), Besag (1974)) is to assume periodicity in the sampling window, equivalently known as wrapping the window onto a torus. On a curved surface (such as the planet Earth) at appropriate scale, then this is a natural definition of edges. So, for example, neighbouring edge sites on the right-hand side are given by the far left border, and vice versa. On a finite lattice this implies that the best estimate for missing edge sites is another 'strip' of sites, which happen to correspond to the border strip furthest away. One of the advantages of this approach is its pleasing symmetries. Every site is used as a neighbour to other sites the same number of times, over the entire lattice. Thus with a simple first-order neighbourhood, every site has four neighbours, and every site is the neighbour to four other sites.

A closely related approach at the other extreme, is to reflect at the boundary, to "mirror" the lattice so that the added "strip" of sites is a replica of the edge sites' neighbours.

Gibbs random fields are the dual representation of Markov random fields, and describe the global, rather than local, relationship between presence/absence values of sites on the lattice.

The mathematically rigorous (measure-theoretic) definition of a *random field* incorporates a sample space, all possible events arising from this sample space, and a measure of probability that may be applied to these events. A *random field* can be defined by the triplet $< \Omega, \mathcal{B}, \mu >$, where Ω is the sample space; $\mathcal{B}$ is the set of all possible events or subsets of Ω (and therefore is a σ-algebra of subsets of Ω, a Borel field); and μ is a probability measure defined on $\mathcal{B}$. Here *events* are configurations of any portions of the lattice within the sample space Ω.

A *Gibbs random field* (GRF) attributes to lattice values a probability measure of a special form, derived from Gibbs distributions of statistical physics. Statistical physics provides the bridge between macro-scale modelling provided by thermodynamics and micro-scale modelling of mechanics (Guénault 1995). In contrast to the full data required when considering data at the micro-scale, or the summaries of data provided by thermodynamics, statistical physics is concerned with describing the statistical distribution of values. Physical

quantities of interest include: pressure, temperature, density, concentration, energy, entropy and free energy. Statistical physics permits absolute values for these quantities to be derived, given relationships described by thermodynamics.

The underlying premise is that all accessible micro-scale states are equally probable. This leads to the need to average over all states to obtain 'ensemble averages', which in standard statistical language are expectations, with respect to a probability density. This has been termed a 'confession of ignorance' over all possible states (Guénault 1995, page 4), reminiscent of a justification often used for explaining use of a Bayesian non-informative prior. The density is the Gibbs distribution, first introduced to describe macroscopic properties of a physical system with Hamiltonian (energy) $\mathcal{H}$. The energy of an open thermodynamical system heated by a reservoir of temperature $\mathcal{T}$ is modelled (Toda, Kubo & Saitô 1983, for example) by a Gibbs distribution:

$$p(z) = \frac{\exp\{-\beta\mathcal{H}\}}{c(\beta;\mathcal{H}(z))} \tag{4.15}$$

where z is the state or configuration of the system, β is a thermodynamic parameter independent of the system, and $c(\cdot)$ is the normalization constant (NC) obtained by integrating over the sample space to ensure the density is proper.

$$c(\beta) = \sum_{z\in\Omega} \exp\{-\beta\mathcal{H}(z)\}. \tag{4.16}$$

To statisticians, this distribution is in fact a member of the familiar exponential family.

The form of the energy function $\mathcal{H}$ defines Gibbs random fields with respect to a neighbourhood system Θ (and neighbourhood basis $\mathcal{N}$). The energy function is expressed as a linear function of functions defined on cliques in $\mathcal{C}$:

$$\mathcal{H}(z) = \sum_{C\in\mathcal{C}} \mathcal{H}_C(z). \tag{4.17}$$

More generally, this can be expressed with respect to functions defined on clique equivalence classes:

$$\mathcal{H}(z) = \sum_{k=0}^{K} \mathcal{H}_k(z) \tag{4.18}$$

where each $\mathcal{H}_k$ corresponds to the kth class of cliques $\mathcal{C}_k(\mathcal{N})$. If only single-site cliques or pairwise cliques are to be considered, then the energy function $\mathcal{H}$ can be written as the sum of two components, corresponding to single-site and pairwise cliques. For example, the pairwise interaction models of Chen (1988) based on those of Besag (1974) are defined by

$$\mathcal{H}(z) = \sum_i U(z_i) + \sum_i \sum_{j\in\mathcal{N}(i)} V(z_i, z_j). \tag{4.19}$$

This is a special case of equation (4.18), with only two classes of cliques $K = 2$.

Furthermore, with presence/absence data the potential functions $\mathcal{H}_k$ can be expressed in terms of the amount of clustering of presence. Count the presences on a sublattice L using

$$\mathcal{H}_0(z) = \sum_{i\in L} I[z_i = 1] \tag{4.20}$$

and let

$$\mathcal{H}_k(z) = \sum_{C \in C_k(\mathcal{N})} I[C \cap L \neq \phi \quad \text{and} \quad z_i = 1 \quad \forall i \in C] \tag{4.21}$$

count the number of k-order cliques which overlap L and contain all presences. This describes the density of small groups of sites having the same spatial relationship (cliques). This corresponds to the global formulation of the autologistic model.

Alternatively the parameters can reflect the clique structure (*e.g.* Possolo (1986*b*)):

$$\beta\mathcal{H}(z) = \sum_{k=0}^{K} \beta_k \mathcal{H}_k(z) \quad \text{or} \quad \beta\mathcal{H}(z) = \sum_{C \in \mathcal{C}} \beta_C \mathcal{H}_C(z) \tag{4.22}$$

with for instance $\mathcal{H}_C(z) = \prod_{i \in C} z_i$ for $\{0, 1\}$ values, or $\mathcal{H}_C(z) = I[z_i = 1 \quad \forall i \in C]$ in general. The model can be simplified by reducing the many parameters (one for each clique of every size and shape), *e.g.* by imposing stationarity on the parameters:

$$\beta_C = \beta_C^* \tag{4.23}$$

for all pairs of cliques C and C^* which are translations of each other. On a regular lattice, cliques have a finite number of easily-defined shapes, so they can be partitioned into equivalence classes based on shape and orientation. Further simplification can be achieved by imposing stationarity on parameters, where cliques C and C^* are related in other symmetric ways.

Another less common alternative is to impose stationarity on the random field directly, rather than on the parameters (Del Grosso 1974)

$$\mathrm{E}[\,\mathcal{H}_C(z)\,] = \mathrm{E}[\,\mathcal{H}_C^*(z)\,] \tag{4.24}$$

whenever cliques C and C^* are equivalent.

The difficulty with discrete GRF models is the normalizing constant (NC) whose evaluation involves summation over the entire sample space. This makes infeasible demands on computing time and space if the NC is found directly, so methods such as Maximum Likelihood and Method-of-Moments are not appropriate. Hence the discovery of an equivalent and corresponding conditional model (Markov random field model) which did not involve this normalizing constant allowed other avenues of estimation to be explored.

4.2.5 Markov property

The Markov property is a defining property of binary Markov random field models which are used to model the underlying spatio-temporal process used in the second half of the thesis (from Chapter 5 onwards).

Central to the concept of the Markov property used in Nearest Neighbour models is the requirement that the conditional distribution of a random variable given several others is dependent only on its nearest neighbours. This is an extension of time series auto-regressive models (*e.g.* Box & Jenkins (1970)). A random field satisfies the *Markov* property if the probability distribution is such that the probability distribution of the presence or absence at a central site conditional on the rest of the lattice is equivalent to that conditional only on the site's neighbours:

$$p(z_I = z \,|\, z_J = \eta) = p(z_I = z \,|\, z_{\mathcal{N}(I)} = \eta) \tag{4.25}$$

for all sublattices $I, J \subset \mathcal{L}$ where the neighbourhood of set I lies within J that is $I \cup \mathcal{N}(I) \subset J$ and the range is defined as $\forall z \in \Omega(I), \quad \eta \in \Omega(J)$ with respect to neighbourhood basis $\mathcal{N}$.

A *discrete Markov random field* (MRF) has a probability measure which satisfies three conditions of positivity, Markovity and homogeneity.

Positivity All configurations of the lattice are possible.

$$p(z_I) > 0 \quad \text{for all} \quad z_I \in \Omega(I) \quad \text{and} \quad I \subseteq \mathcal{L} \tag{4.26}$$

and therefore $0 < p(z_I \mid z_{\mathcal{N}(I)}) < 1$

Markovity The Markov property is defined in Equation 4.25 above.

Homogeneity (Translation invariance) The conditional probability distribution of a site given its neighbours $p(z_i \mid z_{N(i)})$ does not depend on the site i.

$$P\{z_i \mid z_{N(i)}\} = P\{z_j \mid z_{N(j)}\} \qquad \forall i, j \in \mathcal{L} \tag{4.27}$$

or equivalently

$$P\{z_L \mid z_{\mathcal{N}(L)}\} = P\{z_{L+i} \mid z_{\mathcal{N}(L+i)}\} \qquad \forall L \subset \mathcal{L}, i \in \mathcal{L} \tag{4.28}$$

Note that the Markov property is defined in terms of the neighbourhood $\mathcal{N}$ which often is defined to include only nearby sites.

4.2.6 Equivalence between Gibbs and Markov random fields

Due to equivalence between GRFs and MRFs we can express the joint distribution in terms of the equivalence classes of cliques. This result is extremely useful since it simplifies the algorithms for inference for the hierarchical models considered in the thesis.

The Hammersley-Clifford theorem was first proved by the authors it was named after, but apparently never published (Besag 1974, page 197). The theorem was simplified by others including Grimmett (1973), Moussouris (1974), Künsch (1983) and Glötzl & Rauchenschwandtner (1981). It was built on some earlier important results (*e.g.* Dobrushin (1968)). Two important building blocks in this process are the concepts of *Coherence* and *Neighbourhood Influence.*

Coherence ensures that a conditional distribution for some part of the lattice can be obtained from the conditional distribution for a larger portion which contains this part, simply by marginalization. The conditional distributions $\{P(z_I \mid z_{\mathcal{N}(I)}\}$ are *coherent* if they can be obtained from $P\{z_J \mid z_{\mathcal{N}(J)}\}$ whenever $I \cup \mathcal{N}(I) \subset J$. Thus coherent conditional distributions on some part of the lattice are themselves a result of marginalizing at least one joint density defined over the entire lattice. The following theorem tells us that we can reverse this argument and state that coherent conditional distributions may themselves be marginalized further. Dobrushin (1968) also developed this theorem:

Theorem 4.1 *If* $\{P(z_I \mid z_{\mathcal{N}(I)})\}$ *are coherent then there exists at least one translation invariant probability distribution defined on* $\Omega(\mathcal{L})$ *which is also coherent with* $\{P(z_I \mid z_{\mathcal{N}(I)})\}$.

For a given site we will examine the influence of a particular neighbouring site whilst keeping all other neighbours constant. The *influence* of neighbouring site j on central site i can be measured by the "variation distance" between the conditional distributions evaluated at site i when the value of site j is changed:

$$\delta_{ij} = \sup_{\eta \in \Omega(\mathcal{N}(i))} \Big| p(z_i = 1 \,|\, z_j = 0, z_{\mathcal{N}(i)-j} = \eta) - p(z_i = 1 \,|\, z_j = 1, z_{\mathcal{N}(i)-j} = \eta) \Big| \tag{4.29}$$

If the combined influence of neighbouring sites throughout the lattice is small, then (Dobrushin 1968, Föllmer 1982) local and global distributions can be equated. Dobrushin (1968) developed the following theorem:

Theorem 4.2 *If* $\sup_{i \in \mathcal{L}} \{\sum_{j \in \mathcal{L}} \delta_{ij}\} < 1$ *then this is called* weak interaction *and the conditional distributions do combine into a unique joint distribution on* $\Omega(\mathcal{L})$.

Grimmett (1973) showed that weak interaction and coherence hold for Gibbs random fields. Applying the coherence property allows the conditional distribtion of Markov random fields to be derived from the joint Gibbs random field distribution. Refer to these authors for details on the formulation of the argument and for intermediate results. Dubes & Jain (1989) utilized the Hammersley-Clifford theorem to prove the equivalence between a GRF and a MRF in the case of an auto-normal model which only applies to continuous data.

The statement of the theorem by Besag (1974) is simple to understand, and shall be provided here. The reader is referred to this paper for the proof, which incorporates the ideas of coherence and influence. Besag (1974) presents the Hammersley-Clifford theorem by first defining $z^{(i)}$ to be the same as configuration z except that the i component is set to zero, *i.e.* $z_i = 0$. Also define

$$Q(z) = \ln \frac{p(z)}{p(0)} \tag{4.30}$$

where $p(z)$ is the joint distribution of lattice variable z and 0 denotes the configuration of z where all components are 0. Two assumptions made are that $z_i \in \Omega_0$ where Ω_0 is finite, as well as the positivity condition given above in equation (4.26). Then there is a unique expansion of $Q(z)$ of the form

$$\begin{aligned} Q(z) \;=\; & \sum_{i \in \mathcal{L}} z_i \, V_i(z_i) + \sum_{ij \in \mathcal{L}} z_i z_j \, V_{ij}(z_i, z_j) \\ & + \sum_{ijk \in \mathcal{L}} z_i z_j z_k \, V_{ijk}(z_i, z_j, z_k) + \ldots \\ & + z_i z_j z_k \ldots z_L \, V_{ijk,\ldots,L}(z_i, z_j, z_k, \ldots, z_L) \end{aligned} \tag{4.31}$$

where the terms can be expressed in terms of differences between the Q functions, similar to

$$z_i V_i(z_i) \equiv Q(0, \ldots, 0, z_i, 0, \ldots, 0) - Q(0) \tag{4.32}$$

Besag (1974) accredits this theorem to unpublished work by Hammersley & Clifford (1971).

Theorem 4.3 *For any sites indexed by* $1 \leq i < j < \ldots < k \leq L$, *the function* $V_{ij\ldots k}$ *defined in equation (4.31) will be non-zero if and only if the sites* $i, j, \ldots, k$ *form a clique. Providing this condition holds, the* V*-functions can be defined arbitrarily.*

Thus a corollary of the theorem is that a Gibbs random field may be defined by its joint probability density function (*pdf*) or its conditional *pdf* . This is a powerful result considering that it is not always possible to find the analytic form of conditional distributions corresponding to some spatial joint distribution, or vice versa (Cressie 1993). This is what ensures that the auto-models and other similar models are so useful for modelling in a spatial context.

4.2.7 Specific Markov random field models

Specific MRF models provide alternatives to the Autologistic model which is the primary focus of applications to the *dingo* case study from Chapter 5 onwards.

Specific instances of Markov random field models satisfying the general requirements of the Gibbs random fields and the Markov property have been investigated in the literature. A very general class called Pairwise Interaction models consider only 1-site and 2-site cliques. Since computation supporting any inference or simulation increases quickly as the size of the cliques increases, these pairwise interaction models have been the most useful in practice.

Some special cases of the Pairwise Interaction models build on standard statistical distributions: the binomial, logistic, Poisson, and Gaussian. The Derin-Elliott model (Derin & Elliott 1987) is a variation on the auto-binomial and is suitable for categorical data, such as that encountered in image analysis. These are described in the following sections.

Verhagen's model

There are many variations on models tailored to specific applications. Verhagen's (Verhagen 1977, Pickard 1977*b*) model is one such example. It is a physical model which allows for more cliques the pairwise interaction model: horizontal and diagonal cliques (of 2,3, or 4 sites), triplets both horizontal and vertical, and quadruplets horizontal, vertical or diagonal. One parameterization (Possolo 1986*a*) is defined via the local conditional distributions

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} p(z_{st}=1) \\ p(z_{s,t+1}=1 \,|\, z_{st}=1) \\ p(z_{s+1,t+1}=1 \,|\, z_{st}=1, z_{s,t+1}=1, z_{s+1,t}=1) \end{bmatrix} \tag{4.33}$$

Pairwise Interaction General model

Chen (1988) and Dubes & Jain (1989) gave a general anisotropic form of Besag (1974)'s pairwise interaction model for discrete data which is a GRF with an energy function $\mathcal{H}(\cdot)$ of the form shown below. The general form of the joint distribution first given in equation (4.19) can be re-expressed using integer labels for neighbourhood:

$$\mathcal{H}(z) = \sum_{i=1}^{n} U(z_i) + \sum_{i=1}^{n} \sum_{\delta=1}^{\Delta} V(z_i, z_{i:+\delta}) \tag{4.34}$$

where $\Omega_0 = \{0, 1, \ldots, G\}$; $U(\cdot)$ is the potential function for cliques of size 1; and $V(\cdot)$ is the potential function for cliques of size 2. Here V is symmetric, that is $V(z, z*) = V(z*, z)$; and has zero identity, that is $V(z, z) = 0$.

Interpretations for the parameters are shown in Table 4.1.

Any first-order model sets the number of neighbours $|\Delta|$ to 2, so that only the horizontal and vertical spatial dependence parameters are included.

Table 4.1: Interpretation of 1st and 2nd order parameters in general pairwise interaction lattice models.

Parameter	Interpretation
θ_0	abundance or propensity for presence[4]
θ_1	horizontal spatial dependence
θ_2	vertical spatial dependence
θ_3	positive diagonal spatial dependence
θ_4	negative diagonal spatial dependence

Besag (1974) who first defined pairwise interaction models, restricted the form of V to be isotropic with a single factor

$$V(z_i, z_{i:+\delta}) = \beta z_i z_{i:+\delta}. \tag{4.35}$$

We can express the pairwise interaction model in the general notation adopted for Gibbs Random Fields in equation (4.19).

Auto-binomial model

The second-order auto-binomial model for a G-valued discrete map $|\Omega_0| = G < \infty$ allows values to be chosen in neighbourhood according to a binomial distribution:

$$p(z_i \mid z_{N(i)}) = \binom{G-1}{z_i} \pi^{z_i} (1-\pi)^{G-1-z_i} \tag{4.36}$$

where $\binom{a}{b}$ is the usual combinatorics function and the probability of success is given by

$$\pi = \frac{\exp(1-\mathcal{H})}{1+\exp(-\mathcal{H})} \tag{4.37}$$

and the energy function is

$$\mathcal{H} = \theta_0 + \sum_{\delta} \theta_\delta (z_{i:+\delta} + z_{i:-\delta}). \tag{4.38}$$

In the notation of a general pairwise interaction model, the joint probability distribution is:

$$\begin{aligned} U(z_i) &= \theta_0 z_i - \ln \binom{G-1}{z_i} \\ V(z_i, z_{i:+\delta}) &= \theta_\delta \, z_i z_{i:+\delta} \\ |\Delta| &= 4. \end{aligned} \tag{4.39}$$

Autologistic model

The autologistic model is the auto-binomial model with $G = 2$ in equations (4.36)–(4.38). In the notation of a general pairwise interaction model, the joint probability distribution is:

$$\begin{aligned} U(z_i) &= \theta_0 z_i \\ V(z_i, z_{i:+\delta}) &= \theta_\delta \, z_i z_{i:+\delta}. \end{aligned} \tag{4.40}$$

An in-depth treatment of the autologistic model is given in Section 4.3. This model is put to practical use in the *dingo* case study in Chapters 5–7.

Derin-Elliot model

The Derin-Elliot model is an extension of the simple Ising model to a finite discrete valued range set. It was first proposed by Derin & Elliott (1987) in the context of image analysis. It is not strictly an "auto-model" as defined by Besag (1974), which assumes that the spatial dependence measured by $\theta_\delta \equiv \beta$ is constant over all directions in neighbourhood. Instead it is defined by the difference between a spatial dependence parameter θ_δ and the similarity between the sites.

$$\begin{aligned} |\Omega_0| &= G < \infty \\ U(z_i) &= \theta_0(z_i) \\ V(z_i, z_{i:+\delta}) &= \theta_\delta - \mathrm{I}[\, z_i, z_{i:+\delta}\,] \end{aligned} \tag{4.41}$$

where the value of θ_0 is different for each possible value of z_i; and the indicator function $I(a,b) = 1$ if $a = b$; and $= 0$ otherwise.

Auto-Poisson model

The Auto-Poisson model is an extension of the auto-binomial model to an *infinite* discrete valued sample space:

$$\begin{aligned} \Omega_0 &= \{0\} \bigcup \Re^+ \\ U(z_i) &= \theta_0 z_i + \ln z_i! \\ V(z_i, z_{i:+\delta}) &= \theta_\delta\, z_i z_{i:+\delta}. \end{aligned} \tag{4.42}$$

Auto-Gaussian model

The Auto-Gaussian model is an auto model defined on an *infinite* real valued sample space $\Omega_0 = \Re$. The joint probability density is defined by

$$p(z) = \frac{1}{\sqrt{2\pi\sigma^2}}^{L} |\Theta|^{\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(z-\mu)^\top \Theta (z-\mu)\right\} \tag{4.43}$$

where μ is a vector of means, Θ is a positive-definite matrix with $diag(\Theta)$ all ones and off-diagonal elements θ_{ij}. The local conditional formulation is given by

$$\mathrm{E}[\, z_i \,|\, z_{-i}\,] = \mu_i + \sum_j \theta_{ij}(z_j - \mu_j) \tag{4.44}$$

$$\mathrm{Var}[\, z_i \,|\, z_{-i}\,] = \sigma^2 \tag{4.45}$$

Besag (1974) points out that these models, also known as Conditionally AutoRegressive models are different to the Simultaneous AutoRegressive models, as defined in Cressie (1993) for example. Weir & Pettitt (1999) base their auto-probit model on an auto-Gaussian model on a rectangular lattice with $\theta_{ij} = -\alpha$ when i, j are horizontal neighbours $\theta_{ij} = -\beta$ when i, j are vertical neighbours and $\theta_{ij} = 0$ otherwise.

4.3 Anisotropic Autologistic model

This section details an important binary Markov Random field model, the autologistic distribution, which was found useful for application to spatial statistics applications, such as the *dingo* case study (Chapters 5 and 7).

The autologistic was one of several 'auto' models, including the auto-Gaussian and auto-Poisson, which were introduced by Besag (1974). The prefix denotes that these models describe responses autocorrelated in space, analogous to autoregressive models in time series.

The autologistic model is a reparameterization of the Ising model (Ising 1925, Kindermann & Snell 1980), a physical model which in various guises has been used to describe spatial processes in the statistical physics literature.

In the lattice gas context, a variant of the Ising model is expressed in terms of the number of gas particles and the density of these particles (Domb & Green 1972*a*). A physical interpretation of application of Ising's model to the *dingo* case study could be made by analogy to the lattice gas context. The number of dingoes (gas molecules) in a home range (volume), would be distributed with a certain density, and with a certain predisposition to being close to one another (interaction energy). A certain amount of energy would be required for a dingo presence.

Although this thesis focuses on the autologistic parameterization, much information on its theoretical properties is available from the literature on the Ising model. A discussion of parameterization issues on selection of the Ising or autologistic version is given in Sections 4.6, 4.3.2 and 4.3.3.

I begin with a definition of the simple Ising model since it historically predates the autologistic and its one-parameter form is a simple one often used in practice. The autologistic/Ising model is essentially a model for an infinite lattice, and this assumption needs to be taken into account for practical applications. After definition of the basic model in Section 4.3.2 I discuss the issue of finiteness and boundary effects (Section 4.2.3). Section 4.3.4 then describes an important feature of the model called *phase transition*, which occurs at critical values in the parameter space.

> *The Ising model is a crude attempt to simulate the structure of a physical ferromagnetic substance. Its main virtue lies in the fact that a two-dimensional Ising model yields to an exact treatment in statistical mechanics. It is the only non-trivial example of a phase transition that can be worked out with mathematical rigour. (Huang 1987)*

A more indepth exposition of the theoretical results available for the Ising model is given in Section 4.3.5. Unfortunately the literature provides complete results only for the one parameter Ising version. Results for models with more parameters have been more difficult to attain and are still the subject of current research in the statistical physics literature.

> *Since Onsager's solution of the two-dimensional Ising model in zero field, numerous man-hours of effort and many a good researcher falling foul to the dreaded 'Ising's disease' have failed to yield exact solutions to the two-dimensional model in non-zero field and the three-dimensional model, even in zero field. (Baker & Kawashima 1996, page 135)*

Related models arising in image analysis applications are indicated in Section 4.2.7.

Statistical inference for the Ising/autologistic model is hampered by the intractable normalization constant. Analysis therefore generally proceeds via simulation-based methods. In fact the Markov chain Monte Carlo (MCMC) method was first devised by Metropolis, Rosenbluth, Rosenbluth, Teller & Teller (1953) in order to simulate from an Ising model. The MCMC method and variations for simulating from these models are outlined in Section 4.4. Proper implementation of these methods relies on estimating MCMC standard

error (Section 4.4.6) and monitoring MCMC diagnostics (Section 4.4.7). Issues important to practical application of these methods are addressed in Section 4.4.

4.3.1 Definition of Ising Model

The simple Ising model (with no external magnetic field) has only one parameter β measuring similarity between adjacent values. The observation at each lattice site y_i may take on values $+1$ or or -1. The probability density takes the form

$$p(y|\beta) = \frac{h(y,\beta)}{c(\beta)} \tag{4.46}$$

where y refers to the multivariate lattice state[5] $\{y_i : i \in \mathcal{L}\}$; the numerator $h(\cdot)$ is a well defined unnormalized function of y and parameter β; and $c(\beta)$ is the normalization constant, also known as a partition function in statistical physics, which simply integrates h over parameter space Ω. The logarithm of the unnormalized density is given by

$$\ln h(y,\beta) = \beta \sum_{i \sim j} y_i y_j, \qquad y_i \in \{-1,+1\}, \tag{4.47}$$

The sums are over pairs of neighbouring sites i and j, denoted $i \sim j$.

A short digression here allows us to draw parallels between concepts important to statistical physics applications and concepts important to use of the autologistic model in spatial statistics. The benefit of this is that results in statistical physics for the Ising model cannot always directly be translated into a general spatial statistics context (due to the parameterization), although in specific situations this may be possible. These physical concepts include: temperature, energy, spatial interaction, magnetization, specific heat, and Gibbs distributions. Consider the Ising model for ferromagnetism which has Hamiltonian

$$\mathcal{H} = -J \sum_{ij} y_i y_j - H \sum_i y_i \tag{4.48}$$

where y_i is the spin of the atom at site i and takes on values of $+1$ or -1; J is the interaction; and H is magnetic field. Implicitly this corresponds to a canonical ensemble with density given by the Gibbs distribution of equation 4.15 and constant β. Thermodynamic quantities are also pre-defined and involve the normalization constant $c(\cdot)$. The parameter β is factored according to

$$\beta = \frac{J}{k_B \mathcal{T}}, \tag{4.49}$$

where the factor $\frac{1}{k_B \mathcal{T}}$ is independent of the lattice material and incorporates Boltzmann's constant $k_B = 1.38 \times 10^{-16}$ erg/deg and temperature $\mathcal{T}$. The temperature $\mathcal{T}$ can be useful in a non-physical context; simulated annealing (Geman & Geman 1984) controls simulation of the Ising model near phase transition via temperature.

Some of the more basic thermodynamic quantities are related to familiar statistical objects. Free energy is related to the log normalization constant

$$F = -kT \ln c(\beta, H); \tag{4.50}$$

[5] also known as an ensemble or configuration in some statistical physics contexts, and as a map or colouring in image analysis applications

internal energy to the derivative of the log NC with respect to β

$$U = -\frac{\partial}{\partial \beta} \ln c(\beta, H); \tag{4.51}$$

and specific heat to the second derivative

$$C_H = k\beta^2 - \frac{\partial}{\partial \beta^2} \ln c(\beta, H). \tag{4.52}$$

By definition, magnetization corresponds to

$$M = \sum_i \mathrm{E}[y_i], \tag{4.53}$$

the marginal total of Y. This can be interpreted as the overall density of presences observed on the lattice. This density can also be obtained by differentiating the log NC:

$$M = \frac{1}{\beta} \frac{\partial}{\partial H} (\ln c) \tag{4.54}$$

These two ways of expressing the density form the basis of the path integral method for estimating the NC, investigated later in Chapter 6.

4.3.2 Definition of Autologistic Model

Under the transformation $y_i = 2z_i - 1$ (assuming an infinite lattice with torus boundaries) this expression yields the isotropic autologistic model[6] with one parameter:

$$h(z, \theta) = \exp\{-\theta \sum_i z_i + \theta \sum_{i \sim j} z_i z_j\}, \qquad z_i \in \{0, 1\} \tag{4.55}$$

where z_i may take on values 0 or 1, and is an observation on lattice $\mathcal{L}$, at site i, $i \in \{1, \ldots, L\}$, and $c(\theta) = c(\beta)/2$ and $\theta \equiv 4\beta$. With the autologistic parameterization a natural progression to two parameters allows θ to be partitioned into θ_0, θ_1.

$$h(z, \theta) = \exp\{\theta_0 \sum_i z_i + \theta_1 \sum_{i \sim j} z_i z_j\}, \qquad z_i \in \{0, 1\} \tag{4.56}$$

Watson (1972) noted that boundary effects on a finite lattice occur and so this relationship needs to be adjusted in the finite context.

The simple Ising model applies to the situation when $H = 0$ and so $M = 0$ and the marginal average probability of presence over the entire lattice is one half[7]. Thus in the autologistic parameterization this occurs when $\theta_0 = -\theta_1$. This is a reasonable assumption for some applications of the Ising model, such as ferromagnetism, crystallography or lattice gases; for most spatial statistics applications it is not. These include examples given in Section 2.3.

The more general parameterization of the 2-parameter autologistic (equation 4.56) allows unequal proportions of presence/absence over the entire lattice, by permitting θ_0 and θ_1 to vary independently, thereby allowing the marginal average probability of presence over

[6] similar to the structure of the Lattice Gas model of statistical physics although asymptotics differ due to the difference in scales (Binder & Heermann 1997)

[7] there is 'no external field', that is, the overall magnetization is zero

the entire lattice to take on any value. An additional parameter included in the simple Ising model (Kindermann & Snell 1980, Binder & Heermann 1988, for example) to represent an 'external field' gives the same result:

$$\begin{aligned} h(y|\beta) &= \exp\left\{\frac{1}{k_B\mathcal{T}}\mathcal{H}\right\} \\ \mathcal{H} &= J\sum_{i\sim j} y_i y_j + H\sum_i y_i, \qquad y_i \in \{-1,+1\}. \end{aligned} \tag{4.57}$$

where $\mathcal{H}$ is the Hamiltonian, H is named the external field, and $\beta = J/k_B\mathcal{T}$, $\beta_0 = H/k_B\mathcal{T}$. Parameter β_0 is comparable to the 'nugget' effect in geostatistical kriging models (Cressie 1993).

In summary, the Ising parameterization is therefore related to the isotropic two-parameter autologistic via

$$\begin{aligned} p(z|\theta) &= \frac{h(z,\theta)}{c(\theta)} \\ h(z,\theta) &= \exp\theta^{\mathsf{T}}V(z) \\ \theta &= \begin{bmatrix} \theta_0 & \theta_1 \end{bmatrix}^{\mathsf{T}} \\ &\equiv \begin{bmatrix} \beta(H-4J) & 4\beta J \end{bmatrix}^{\mathsf{T}} \\ V(z) &= \begin{bmatrix} \sum_{i=1}^{L} z_i & \sum_{i\sim j} z_i z_j \end{bmatrix}^{\mathsf{T}} \\ &\equiv \begin{bmatrix} \frac{1}{2}(L+\sum_i y_i) & \frac{1}{4}\sum_{i\sim j} y_i y_j + \sum_i y_i + \frac{L}{2} \end{bmatrix} \end{aligned} \tag{4.58}$$

$$\ln c(z,\theta) \equiv \ln c(y,\beta) - L\beta(2J-H) \tag{4.59}$$

These relationships to the Ising model are provided here to assist with reading the statistical physics literature and to assist with interpretation of theoretical results provided within this chapter.

4.3.3 Isotropy

The 2-parameter autologistic or Ising models each assume that autocorrelation is isotropic, *i.e.* autocorrelation is the same in every direction from any given site on the lattice. One can decompose the spatial interaction term $\sum_{i\sim j} z_i z_j$ into components for different neighbourhood relationships. This can be rewritten as $\sum_i \sum_{j\in\mathcal{N}(i)} z_i z_j$ to generalize the isotropic neighbourhood operator $\sim$. Here $\mathcal{N}(i)$ denotes the set of sites $j \in \mathcal{L}$ which are neighbours of site i, so $\mathcal{N}(i) = \{j : j \sim i\}$ for an isotropic neighbourhood. Using a geometric approach to neighbourhoods one can define a finite partition of the neighbourhood $\mathcal{N}$ for general site i as

$$\mathcal{N}(i) = \bigcup_{k=1}^{K} \mathcal{N}_k(i) \quad \text{with} \quad \mathcal{N}_k(i) \cap \mathcal{N}_{k^*}(i) = \emptyset, \text{ for } k \neq k^*. \tag{4.60}$$

Splitting the neighbourhood $\mathcal{N}(i)$ and the parameter θ according to this partition achieves an anisotropic parameterization

$$h(z,\theta) = \theta_0 \sum_{i=1}^{n} z_i + \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j\in\mathcal{N}_k(i)} \theta_k z_i z_j \tag{4.61}$$

So for multidimensional $z \in \Omega$ with binary values $z_i \in \{0,1\}$ observed at sites i on lattice L, the $K+1$ parameter autologistic model (with a constant) is defined by equation (4.1) with

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_K \end{bmatrix}, \qquad V(z) = \begin{bmatrix} \sum_{i \in L} z_i \\ \sum_{i \in L} z_i \sum_{j \in \mathcal{N}_1(i)} z_j \\ \vdots \\ \sum_{i \in L} z_i \sum_{j \in \mathcal{N}_K(i)} z_j \end{bmatrix}. \tag{4.62}$$

The parameter θ_0 controls the prevalence of presences on the lattice, and the remaining K autocorrelation parameters $\{\theta_k,\ k = 1, \ldots, K\}$ measure the dependence between pairwise neighbours of different lags and directions. The vector $V(z)$ is the canonical statistic in the exponential family as defined by Cox & Hinkley (1974).

Economies in parameterization can be achieved via symmetry. For example, consider the first order neighbourhood in two dimensions on a square lattice

$$\mathcal{N}(i) = \{i^* + u, i^* + v, i^* - u, i^* - v\}, \qquad u = (1,0), \quad v = (0,1)$$

where i^* is the two-dimensional Cartesian coordinate corresponding to site index i. Positions u and v represent the generic offsets of the first order horizontal (East) and vertical (North) neighbours. Then

$$h(z,\theta) = \theta_0 \sum_i z_i + \sum_i z_i \left(\theta_1 z_{i+u} + \theta_2 z_{i+v} + \theta_3 z_{i-u} + \theta_4 z_{i-v}\right).$$

Invoking symmetry along the East-West and North-South axes via $\theta_1 = \theta_3$ and $\theta_2 = \theta_4$ gives the simpler

$$h(z,\theta) = \theta_0 \sum_i z_i + \sum_i z_i \left(\theta_1 z_{i+u} + \theta_2 z_{i+v}\right). \tag{4.63}$$

An extra factor of two is absorbed into the NC. This equation can be rewritten using the notation of Section 4.2 as

$$h(z,\theta) = \theta_0 \sum_i z_i + \sum_i z_i \left(\theta_1 z_{i:+1} + \theta_2 z_{i:+2}\right).$$

For the autologistic model with the natural statistic given by equation (4.62), the probability distribution of a particular site's value given the configuration at all other sites is:

$$p(z_i | z_{L \setminus i}, \theta) = p(z_i | z_{\mathcal{N}(i)}) = \frac{\exp z_i \{\theta_0 + \sum_{k=1}^K \theta_k \sum_{j \in \mathcal{N}_k(i)} z_j\}}{1 + \exp\{\theta_0 + \sum_{k=1}^K \theta_k \sum_{j \in \mathcal{N}_k(i)} z_j\}} \tag{4.64}$$

Alternatively parameters can be used to describe neighbourhood relationships instead of subscripts for the data z as above in equation (4.55). In the context of spatial linear models, Cressie (1993) writes

$$h(z,\theta) = \sum_i \theta_i z_i + \sum_{1 \le i,j \le n} \theta_{ij} z_i z_j \text{ with } \theta_{ij} = 0 \text{ unless } i \sim j;\ \theta_{ji} = \theta_{ij};\ \theta_{ii} = 0.$$

Expression 4.61 leads naturally into the autologistic model with covariates of Denham & Mengersen (1999), Huffer & Wu (1998) and Preisler (1993). They combine a linear predictor based on covariate x with a simple isotropic autologistic:

$$h(z, \theta, \alpha, x) = \alpha_0 + \sum_{k=1}^{K} \alpha_k x_k + \theta \sum_{1 \le i,j \le n} z_i z_j \tag{4.65}$$

where α_0 is the intercept term, and α_k are coefficients of covariates x_k.

Isotropy in the autologistic model is restricted to two directions for the three-parameter Autologistic distribution used later in the thesis (Chapters 5–7).

4.3.4 Phase Transition

Heuristically it is understandable that once there is a high degree of spatial dependence, if the state of just one site changes, then the 'domino effect' on its neighbours and their neighbours and so on, could easily mean that large changes in the overall lattice are just as likely as the current state, even states that substantially alter the balance between presence or absence. Essentially in some *critical* regions of the parameter space $\theta \in \Theta_{\text{crit}}$, representing a high degree of spatial dependence or interaction, the NC can develop singularities[8]. This in turn can affect asymptotics. So in practice, large patches can appear consisting of presence and absence. This is because the sampling distribution of the mean of the lattice has two maxima (*e.g.* Kindermann & Snell (1980)).

These critical values are not reported specifically in the statistical literature, despite their impact on modelling. The critical values for β in the single parameter Ising model on an infinite lattice satisfy (Huang 1987):

$$\frac{J}{kT} = \beta > \beta_{\text{C}} = \frac{1}{2}\ln(1+\sqrt{2}) = 0.4406868. \tag{4.66}$$

Two sets of lattice values simulated under the same parameter β could thus be entirely different in character and in sampling mean.

To date, no exact results for evaluation of the NC or the critical point have been obtained for versions of the Ising model which do not have zero external field (Thompson 1988, Baker & Kawashima 1996). However one can estimate the corresponding autologistic critical parameter value, keeping in mind that it is only strictly correct when $\theta_0 = -\theta_1 = -\theta_2 = \theta$:

$$\theta > \theta_{\text{C}} = 4\beta_{\text{C}} = 1.762747. \tag{4.67}$$

Experimental results (Chapter 7) on a finite lattice comprising 1000 sites confirm that when the spatial or temporal dependence parameters θ_1, θ_2 values exceed 2.0 simulations can be extremely variable and estimates based on these numerically unstable. In these situations, boundaries can influence the entire lattice, and long range dependence is not negligible.

According to a "History of the Ising model" presented in Brush (1967), phase transition was precisely the feature that the creators Lenz and student Ising (Lenz 1920, Ising 1925) were aiming for when constructing the two-dimensional Ising model. However this property was not confirmed until a series of papers with Peierls (1936) and Onsager (1944) providing vital breakthroughs. (See Kindermann & Snell (1980) for an introductory exposition.)

[8]When the Ising model is applied to ferromagnetism, *phase transition* occurs at *the Curie temperature*, and below these temperatures the ferromagnet may 'spontaneously' magnetise in either direction along its principal axis.

Older work (before the twentieth century) to understand phase transitions included the Weiss mean field theory of magnetism and the theory of liquid-vapour transition by Van der Wal (Burley 1972, for details). The advantages of these theories are their simplicity, closed form expression, qualitative accuracy, and fair quantitative accuracy in the parameter space away from the phase transition.

The region of the parameter space leading to phase transitions greatly affects inference when high levels of spatio-temporal dependence are considered. We see later that many existing methods of inference fail at these levels of dependence. The critical values given here help give a practical guide as to when this failure might be expected. Practical results in Chapters 6 and 7 confirm that these critical values are useful.

4.3.5 Theoretical results for the Ising model

Theoretical results for the Ising model provide important information for explaining the properties and potential applications of both the Ising and the autologistic models. In particular the phenomenon of phase transition affects inference for parameters. Furthermore, work by statistical physicists designed to explore the properties of the Ising model, have as their by-product produced results for the log normalization constant (NC). These results are therefore important to the examination of methods for estimating log NC ratios in Chapter 6.

The two-dimensional isotropic Ising model, as defined by equation (4.47), is important in statistical physics due to its simplicity and availability of exact results. In the statistical physics context, 'results' include expressions for the normalization constant and for critical parameters, as well as for other thermodynamic parameters of interest, such as internal energy, specific heat, etc (see page 81). These critical parameters correspond to phase transitions observed in reality, such as the well known gas-liquid and liquid-solid transitions of matter, and the spontaneous magnetization of a ferromagnetic below certain temperatures. Ising (1925) himself found that the one-dimensional Ising model did not exhibit phase transitions, and erroneously concluded that this must also be the case for the two-dimensional version. Results for the critical parameters and a closed expression for the normalization constant were later derived, however, for the two-dimensional isotropic version of the Ising model with no external field. They were arrived at gradually via several approximations (Bragg & Williams 1934, Bethe 1935, Peierls 1936) before the exact result was announced by Onsager (1944).

The Bragg-William approximation (Bragg & Williams 1934) hinges on two important quantities corresponding to local and global marginal probabilities of presence in statistical language. The basic assumption is that the local marginal probability of presence for a pair of sites, given its neighbourhood, can be derived from the global marginal probability of presence for any site, regardless of neighbourhood. Discussion of this approximation highlights the capacity for the Ising/autologistic model to capture long-range dependence (LRD) in two or three dimensions. Long range dependence is a current topic for research in statistics, where activity is focussed on finding time series models capable of exhibiting this trait, corresponding to phase transitions in the statistical physics literature. This assumption simplifies and therefore does not accurately reflect the relationship between values at two neighbouring sites. The specific heat (or Hessian matrix of the normalization constant) disappears for certain regions of the parameter space.

The Bethe-Peierls approach (Bethe 1935, Peierls 1936) improves the representation of this relationship by considering sublattices, comprising a site and its neighbours, to be

immersed within the overall lattice. The assumption is that the rest of the lattice impacts on a sublattice in a way that depends on a function of the spatial interaction parameter called *fugacity*. The specific heat (or Hessian matrix) near the critical point is modelled better with the Bethe-Peierls than with the Bragg-Williams approximation yet still does not capture the true property of the Ising/autologistic model. This is reminiscent of nearest neighbour models where a covariate averaging response within the neighbourhood is used to represent the impact of neighbourhoods. The coefficient of this covariate is estimated by considering all sites over the lattice and in some ways therefore parallels the fugacity parameter of the Bethe-Peierls approximation.

Neither of the earlier approximations, Bragg-Williams or Bethe-Peierls, were sufficiently accurate to capture the nature of the phase transition, instead confirming predictions of classical theory. Onsager's solution was thus important in establishing that the simple 2D Ising model had a phase transition, characterized by a divergence in the specific heat that was symmetric on the logarithmic scale. Recall (page 81) that specific heat is related to the second derivative of the log NC with respect to a component of the interaction parameter. This result opposed classical thermodynamic theory, and provided researchers with a theoretical framework in which to investigate phase transitions. It later provided a basis for explaining and predicting liquid gas transitions (Yang 1972).

Since then other mathematical approaches have simplified, extended, and confirmed Onsager's result (Yang & Lee 1952). Other approximations have been devised and tested for accuracy against Onsager's result, and then been used to predict phase transitions in physical contexts using models other than the two-dimensional Ising model. Several approximations have been based on series expansions which refer to "lattice constants" or the geometry of all different types of graph that can be constructed from bonds on a lattice (Domb & Green 1974). The series are termed perturbation expansions in the spatial interaction parameter. With low interaction the series are smooth analytic functions, but gradually approach functions with discontinuities at phase transition for higher levels of interaction. These approximations include: a method based on free random walks paralleling quantum theory (Wortis 1974); and the Padé approximant (Baker 1961, Gaunt & Guttmann 1974) which is essentially a power series expansion near the critical point. Patterns of phase transition in general are thought to be governed by the *Universality principle* which states that it is the lattice dimension and value at individual sites, rather than the lattice geometry, which affects phase transition.

Renormalization theory, simultaneously discovered by several researchers (Domb & Green 1976, for references), provided a breakthrough. This strategy, including two expansions called the ϵ and $1/n$ expansions, was adopted from quantum theory and makes reference to the Central Limit Theorem.

> *One of the most beautiful aspects of the renormalization group approach is the wealth of useful physical information which can be obtained from multiple fixed points, flow diagrams and crossover effects. (Domb & Green 1976).*

Feynman diagrams are a powerful graphical method which facilitate analysis of physical models in the area of quantum mechanics. These could be applied to the critical point once it was realised that the suitable value of the expansion parameter was $\epsilon = 4 - d$, where d is the dimension of the lattice. The renormalization approach proceeds by manipulating exact analytic expansions with possibly infinite coefficients until the singularities disappear. The quantum approach also highlights that the critical singularities arise from the fact that "fluctuations at the critical point do not have a characteristic length" (Wilson 1976). This

can be interpreted to mean that the variability at the critical point is due to several sources, including both long range lattice correlation and spacing between sites. To cope with the multiple sources of error, successful strategies have broken the problem down into sub-problems (Niemeijer & van Leeuwen 1976).

Later work has focussed on determining whether a similar phase transition occurs for similar models in higher dimensions. This has given rise to a large body of research on the "hyperscaling hypothesis".

Although approximations were devised by statistical physicists to inspect behaviour of models (such as the Ising) near the critical region, these approximations are of interest in this thesis since they relate to the log NC. The NC is important for a fully Bayesian hierarchical model applied to the *dingo* case study in Chapter 7. Thus it is of interest to provide details of the simplest approximations: Bragg-Williams and Bethe-Peierls. The exact method of Onsager, subsequent improvements, and the renormalization group approach involves quite complex mathematics. The reader is invited to consult any text on statistical physics (Thompson 1988, Domb & Green 1972*a*, Kindermann & Snell 1980, Huang 1987) using some of the notes in this section to translate concepts back into the statistical arena. The hyperscaling hypothesis is briefly discussed last in this section since this indicates how current research may provide new information about the Ising model. Finally, the relationship to Gaussian models (Section 4.3.5) is of interest since it provides an alternative to the Ising/autologistic in some situations.

I follow aspects of the mathematical presentation in Thompson (1988), the elementary of Huang (1987), and the physics of Temperley (1972) and Domb & Green (1972*c*). I have translated results to the statistical context where possible, however the thermodynamic variable temperature often has no equivalent (that I am aware of) in this context.

Dual representation

The statistical physics expression of the Ising model on a lattice where each site has 4 neighbours was given in equation (4.57). These planar results can be generalized to lattices of different geometries, *e.g.* hexagonal, triangular in 2D; cubic in 3D; polymer walks (Domb & Green 1972*c*). For simplicity I focus on the square lattices which correspond to motivating applications of Chapter 2. Where possible I translate results presented in the statistical physics literature for the Ising model to the autologistic using the relationship given in Section 4.3.2. This translation only becomes difficult when prevalence and spatial interaction parameters are factored by the constant $k_B T$, as given by equation 4.49.

Define

$$\begin{aligned} V_+ &= \sum_i I[z_i = 1] = \sum_i z_i \\ V_{++} &= \sum_{i \sim j} I[z_i = 1, z_j = 1] = \sum_{i \sim j} z_i z_j \end{aligned} \tag{4.68}$$

where $I[\cdot]$ is the indicator function with value one when true and zero when false. Thus V_+ and V_{++} correspond to the canonical statistics of the autologistic[9]Note that $V_+ = \sum_i I[y_i = +1]$ and $V_{++} = \sum_{i \sim j} I[y_i = +1, y_j = +1]$ in the Ising model context.

[9]The lattice gas model is closer to the autologistic in some ways. Important parameters include: V_+ the total number of atoms; V_{++} the total number of nearest neighbour atoms; and $J_0 \equiv 4J$ the interaction energy between a pair of nearest neighbour atoms. The interaction energy between any other atom pairs is assumed negligible. See Syozi (1972) for more details.

Similarly V_-, V_{--}, and V_{+-} can also be expressed in terms of V_+ and V_{++}. First note that by definition $L = V_+ - V_-$. So $V_- = L - V_+$.

Following Peierls (1936), Griffiths (1964) and later Swendson & Wang (1987), one can visualize the lattice in terms of the relationships or bonds between sites, instead of seeing the lattice as a web of sites with presence and absence values. Imagine creating a bond between any site with a presence and each of its neighbours[10]. Then the total number of bonds is $4V_+$. Note that there will be two bonds between every $(++)$ pair, one bond between each $(+-)$ pair and zero bonds between every $(--)$ pair. Thus $4V_+ = 2V_{++} + V_{+-}$ and by symmetry $4V_- = 2V_{--} + V_{-+}$. So

$$V_{+-} = 4V_+ - 2V_{++} \tag{4.69}$$

So the Ising canonical statistics are

$$\sum_{i\sim j} y_i y_j = V_{++} + V_{--} - V_{+-} = 4V_{++} - 8V_+ + 2L \tag{4.70}$$

$$\sum_{i=1}^{L} y_i = V_+ - V_- = 2V_+ - L \tag{4.71}$$

Bragg-Williams approximation

The earliest approximation to the NC was provided by Bragg & Williams (1934) and begins with the concepts of what they term long-range order $\gamma_+^{\mathcal{L}}$ and short-range $\gamma_{++}^{\mathcal{N}}$ order. These measures of order correspond to the statistical measures of global marginal and local conditional probabilities of presence $p(z_i = 1)$ and $p(z_i = 1|z_{\mathcal{N}(i)})$ respectively. They are defined as follows.

Consider all configurations of the lattice as being equilikely[11]. For simplicity we assume that the neighbourhood comprises first order neighbours. Given presence at a particular site then the proportion of first order neighbours also present is $V_{++}/(2L)$; this quantity therefore reflects local conditional probability of presence. In contrast, given that one knows the state of a particular site, then this does not alter the overall proportion of presence V_+/L over the entire lattice; this quantity can therefore reflect the global marginal proportion of presence. These quantities can be translated to the interval $[-1, 1]$ as

$$\gamma_+^{\mathcal{L}} = \frac{V_+}{L} - 1, \qquad -1 \le \gamma_+^{\mathcal{L}} \le +1; \tag{4.72}$$

$$\gamma_{++}^{\mathcal{N}} = \frac{V_{++}}{2L} - 1, \qquad -1 \le \gamma_{++}^{\mathcal{N}} \le +1 \tag{4.73}$$

The approximation used by Bragg & Williams (1934) amounts to assuming that the local conditional probability of presence of a pair can be derived from the global probability of the pair, and that furthermore the latter is given by the product of the probabilities of each individual. That is

$$\left(\frac{V_+}{L}\right)^2 = \frac{V_{++}}{2L}$$
$$\text{or } \gamma_{++}^{\mathcal{N}} \approx \tfrac{1}{2}(\gamma_+^{\mathcal{L}} + 1)^2 - 1 \tag{4.74}$$

[10] Compare this to Kindermann & Snell's (1980) formulation with even bonds and odd bonds, which correspond to the double bonds and single bonds described here.

[11] the principle of ignorance mentioned on page 72

We now follow the derivation of the approximation based on the Ising model, and later translate results back to the autologistic context.

Using (4.74) the Hamiltonian $\mathcal{H}$ of the Ising model becomes

$$\frac{\mathcal{H}(y)}{L} \approx 2J(\gamma_+^{\mathcal{L}})^2 + H\gamma_+^{\mathcal{L}} \tag{4.75}$$

The NC contains a summation over all configurations y. With this approximation, however, the summand only depends on $\gamma_+^{\mathcal{L}}$, which by definition depends on V_+. Now the number of configurations which have the same V_+ is the number of ways of choosing V_+ from L. So

$$c(y,\beta) = \sum_{\gamma_+^{\mathcal{L}}=-1}^{+1} \frac{L!}{[\frac{1}{2}L(1+\gamma_+^{\mathcal{L}})]![\frac{1}{2}L(1-\gamma_+^{\mathcal{L}})]!} \exp\left\{\beta L\left(2J[\gamma_+^{\mathcal{L}}]^2 + H\gamma_+^{\mathcal{L}}\right)\right\} \tag{4.76}$$

Sterling's approximation can be applied: as $L \to \infty$ the log NC approximates the logarithm of the largest term in the summand[12]. Therefore the log NC can be written

$$\frac{1}{L}\ln c(y,\beta) = \beta(2J\overline{\gamma}^2 + H\overline{\gamma}) - \frac{1+\overline{\gamma}}{2}\ln\left(\frac{1+\overline{\gamma}}{2}\right) - \frac{1-\overline{\gamma}}{2}\ln\left(\frac{1-\overline{\gamma}}{2}\right) \tag{4.77}$$

where $\overline{\gamma}$ is the value of $\gamma_+^{\mathcal{L}}$ which maximises the summand of equation (4.76). Thus $\overline{\gamma}$ is the root of the equation

$$\ln\left(\frac{1+\overline{\gamma}}{1-\overline{\gamma}}\right) = 2\beta H + 8\beta J\overline{\gamma} \tag{4.78}$$

and hence

$$\overline{\gamma} = \tanh\left(\beta H + 4J\overline{\gamma}\beta\right). \tag{4.79}$$

In the situation with even marginal probabilities of presence ($H = 0$) there is a simple solution

$$\overline{\gamma} = \tanh\left(4J\overline{\gamma}\beta\right). \tag{4.80}$$

A solution is to compare the curves $f(\gamma)$ given in equation (4.80) and $f(\gamma) = \gamma$. These curves intersect in the following cases

$$\begin{aligned} \overline{\gamma} &= 0 && (4J\beta < 1) \\ \overline{\gamma} &= \begin{cases} \gamma_* \\ -\gamma_* \end{cases} && (4J\beta > 1) \end{aligned} \tag{4.81}$$

for some value γ_* to be evaluated below. When $J > 0$ there exists a critical or Curie Temperature T_c given by

$$\beta_c^{-1} = kT_c = 4J \tag{4.82}$$

satisfying

$$\overline{\gamma} = \begin{cases} 0 & (T > T_c) \\ \pm\gamma_* & (T < T_c) \end{cases} \tag{4.83}$$

Since $\overline{\gamma}$ is the magnetization per site, this makes it clear that for $T < T_c$ the system is a ferromagnet, *i.e.* marginal probabilities of presence and absence are not equal, but for $T > T_c$ the system has no magnetization.

[12]This is the Laplacian approximation used in Lewis & Raftery (1997).

Some approximations can be obtained for γ_* as follows:

$$\begin{aligned} \gamma_* &\approx 1 - 2e^{-2T_c/T} \quad \left(\tfrac{T_c}{T} << 1\right) \\ \gamma_* &\approx \sqrt{3\left(1 - \tfrac{T}{T_c}\right)} \quad \left(0 < 1 - \tfrac{T}{T_c} << 1\right). \end{aligned} \tag{4.84}$$

Thermodynamic applications (Thompson 1988, for example) then focused on evaluating quantities such as the specific heat, magnetization, *etc* at the critical temperature. The evaluation of specific heat at the critical temperature computed using this approximation gives results which are clearly not supported by experimental data, indicating the method's drawbacks.

Now returning to the autologistic parameterization, this gives a critical value of the spatial interaction parameter θ_1 given by

$$\theta_c = 1 \tag{4.85}$$

satisfying

$$\overline{\gamma} = \begin{cases} 0 & (\theta_1 < \theta_c) \\ \pm\gamma_* & (\theta_1 > \theta_c) \end{cases} \tag{4.86}$$

where $\overline{\gamma}$ represents the marginal mean of the configuration z_* giving the largest contribution to the log NC:

$$\overline{\gamma} = 1 - \frac{2}{L}\sum_i z_{*i}. \tag{4.87}$$

So for $\theta_1 < 1$ this marginal mean is zero, however for $\theta_1 > 1$ this is non-zero. Furthermore

$$\begin{aligned} \gamma_* &\approx 1 - 2e^{-2\theta_1} \quad (\theta_1 << 1) \\ \gamma_* &\approx \sqrt{3\left(1 - \theta_1\right)} \quad (0 < 1 - \theta_1 << 1) \end{aligned} \tag{4.88}$$

Thus an approximation to the log NC is possible using a combination of equations (4.88), (4.86), (4.59) and the expansion of NC given in equation (4.77)

$$\frac{1}{L}\ln c(z,\theta) = \beta(2J\overline{\gamma}^2 + H\overline{\gamma}) - \frac{1+\overline{\gamma}}{2}\ln\left(\frac{1+\overline{\gamma}}{2}\right) - \frac{1-\overline{\gamma}}{2}\ln\left(\frac{1-\overline{\gamma}}{2}\right) - \beta(2J - H) \tag{4.89}$$

This approach is an application of the more general "mean-field theory" (Huang 1987, for example). Here one assumes that each site is impacted by a mean field due to all its neighbours. This general formulation utilizes Landau theory to make a simple approximation to the NC by replacing it with the maximum value of the integrand to obtain a saddle-point approximation. Maximum-likelihood which finds the value of the parameter which maximises the integrand of the NC is therefore closely related.

Bethe-Peierls approximation

This approach improves the Bragg-Williams approximation by including the local marginal distribution. The assumption equation (4.74) which ignores local correlation between sites is replaced by finding a more accurate relationship between V_{++} and V_+. This is achieved by moving focus from the entire lattice to a sublattice, comprising a designated site and its nearest neighbours, which can be thought of as being immersed within the global lattice. The next step is to assume that the background affects the sublattice via a transformation of the prevalence parameter $\phi = \exp\{-2\theta_0\}$ in the autologistic or fugacity $\phi = \exp\{2\beta(4J -$

$H)\}$ in the Ising model. A relationship between V_+ and V_{++} is constructed on the sublattice, and then extrapolated to the global lattice. Again, only the case with no external field $H = 0$ is considered.

Again we derive the approximation using the Ising parameterization and then translate results into the autologistic framework.

Let $p(y_i, r)$ denote the probability that the state at site i has value y_i and r neighbours are positive. If $y_i = 1$ then $p(y_i, r)$ counts configurations of the sublattice with r $(++)$ pairs and $4 - r$ $(+-)$ pairs. Conversely if $y_i = -1$ then the sublattice configurations referred to are r $(-+)$ pairs and $4 - r$ $(--)$ pairs. For fixed r there are $\binom{4}{r}$ ways of choosing the positive neighbours. Thus

$$p(y_i = +1, r) = \frac{1}{c(\beta)} \binom{4}{r} e^{\beta J(2r-4)} \phi^r \tag{4.90}$$

$$p(y_i = -1, r) = \frac{1}{c(\beta)} \binom{4}{r} e^{\beta J(4-2r)} \phi^r \tag{4.91}$$

where $c(\beta)$ is the NC and ϕ is the parameter introduced to represent the effect of the background provided by the rest of the lattice. Parameter ϕ is similar to the thermodynamic concept called fugacity. The NC $c(\beta)$ can be determined by applying the law of total probability:

$$\sum_{r=0}^{4} [p(+1, r) + p(-1, r)] = 1 \tag{4.92}$$

so

$$c(\beta) = \sum_{r=0}^{4} \binom{4}{r} \left[(\phi e^{2\beta J})^r e^{-\beta J 4} + (\phi e^{-2\beta J})^r e^{\beta J 4} \right] \tag{4.93}$$

$$= (e^{\beta J} + \phi e^{-\beta J})^4 + (\phi e^{\beta J} + e^{-\beta J})^4. \tag{4.94}$$

By definition one can express the global and local marginal probabilities of presence $\gamma_+^{\mathcal{L}}$ and $\gamma_{++}^{\mathcal{N}}$ in terms of fugacity ϕ.

$$\begin{aligned} \frac{1+\gamma_+^{\mathcal{L}}}{2} \equiv \frac{V_+}{L} &= \textstyle\sum_{r=0}^{4} p(+1, r) &&= \frac{1}{c(\beta)} \left(e^{\beta J} + \phi e^{-\beta J} \right)^4 \\ \frac{1+\gamma_{++}^{\mathcal{N}}}{2} \equiv \frac{V_{++}}{\frac{1}{2}4L} &= \tfrac{1}{4} \textstyle\sum_{r=0}^{4} r p(+1, r) &&= \frac{\phi}{c(\beta)} \left(e^{-\beta J} + \phi e^{\beta J} \right)^3 \end{aligned} \tag{4.95}$$

Assuming that these properties hold throughout the lattice, the expression for the Hamiltonian can be used to compute the NC for the Ising model. However a shortcut to computing the magnetization is available which does not require computation of the NC.

The probabilities in equation (4.90) and equation (4.91) measure the likelihood of each possible value of a central site y_i given the number of neighbours having each value r. Thus

$$\sum_{r=0}^{4} p(y_i = +1, r) \tag{4.96}$$

represents the overall probability of finding a positive value at the central site. In addition, the expression

$$\frac{1}{4} \sum_{r=1}^{4} r \left[p(y_i = +1, r) + p(y_i = -1, r) \right] \tag{4.97}$$

can be interpreted as the expected proportion of positive valued neighbours, or equivalently the probability that a neighbour is positive. In this construction, neither of these probabilities is conditioned on any premise, so they can be equated to achieve a consistent probability of positive values throughout the lattice. (This implicitly considers behaviour of each sublattice separately and ignores the interactions between these.) Equating equation (4.96) and equation (4.97) then determines ϕ

$$\begin{aligned}(e^{-\beta J}+\phi e^{\beta J})^4 &= \frac{\phi}{4}\frac{\partial}{\partial\phi}\left[(e^{-\beta J}+\phi e^{\beta J})^4+(e^{\beta J}+\phi e^{-\beta J})^4\right] \\ &= \phi\left[(e^{-\beta J}+\phi e^{\beta J})^3 e^{\beta J}+(e^{\beta J}+\phi e^{-\beta J})^3 e^{-\beta J}\right] \\ \phi &= \left(\frac{1+\phi e^{2\beta J}}{\phi+e^{2\beta J}}\right)^3 . \end{aligned} \tag{4.98}$$

Solving for ϕ and using equation (4.95) gives

$$\overline{\gamma_+^{\mathcal{L}}} = \frac{\phi^a-1}{\phi^a+1}, \qquad a=\frac{4}{3} \tag{4.99}$$

$$\overline{\gamma_{++}^{\mathcal{N}}} = \frac{2\phi^2}{(1+\phi e^{-2\beta J})(1+\phi^a)}-1 \tag{4.100}$$

The critical temperature can be computed as

$$T_c = \frac{1}{a}\frac{2J}{\ln 2} \tag{4.101}$$

The problem with the specific heat function does not occur in contrast to the Bragg-Williams result.

Recasting results into the autologistic framework by applying equation (4.59) to equation (4.101), we have

$$\theta_c = 2\ln 2 = 1.386294. \tag{4.102}$$

This obviously differs from the value obtained with the Bragg-Williams approximation. It corresponds more closely to the behaviour observed with the simulations in Chapter 7, where models with spatial or temporal interaction parameters greater than 1.25, *i.e.* $\theta_1 > 1.25, \theta_2 > 1.25$ caused divergences in the numerical algorithms used to estimate log NC ratios comparing two θ's.

Onsager's method and other exact methods

The exact method evolved from the work of Kramers & Wannier (1941), and Onsager (1944). The work of Yang & Lee (1952) was seminal in clarifying and simplifying this work. A number of different mathematical approaches have verified the initial results.

Kramers & Wannier (1941) found that the normalization constant (NC) of a square lattice at low temperature (high θ) could be related to the corresponding NC at high temperature (low θ). Onsager (1944) found a topological interpretation for this result and extended it so that the NC of any planar lattice could be related to that of its dual. The presence/absence data could be equivalently considered as a set of closed loops of bonds on the dual lattice, each loop dividing presences from absences. This allowed critical parameters to be obtained precisely. This work proceeded by defining two operators; V_2 described interactions between all site variables within a row of m sites and V_1 described

interactions betweeen all sites in two neighbouring rows. Then an exact relationship was derived between the NC of a 2D lattice (using torus geometry) and the trace of the nth power of operator $V = V_2 \cdot V_1$. For m, n large this reduces to a problem of computing the largest eigenvalue of V. The Lie Algebra generated by V_1 and V_2 (via Fourier transforms of these operators) could be written as products of operators represented by 2×2 matrices. So the problem reduced to finding the maximum product of eigenvalues of these 2×2 matrices.

An underlying reason for the Lie Algebra step was due to the isomorphism of spinor and rotation groups. This simplified the work of Onsager (1944) and allowed computation of correlation functions $\mathrm{E}[\, y_{ij} y_{i+k,j+l}\,]$. Closed forms become difficult as k, l grow large but can be written down as determinants of the Toeplitz type which can be asymptotically evaluated, to give limiting long range correlations. The one-step row or column correlations $\mathrm{E}[\, y_{ij} y_{i+1,j}\,]$ and $\mathrm{E}[\, y_{ij} y_{i,j+1}\,]$ were derived by Onsager (1944).

The long-range correlation is zero above the critical temperature but is finite below. Finite positive long range correlation implies that any two sites are likely to have the same value, leading to the phenomenon where a spontaneous excess of presences/absences can occur. The limiting long range correlation is equal to the magnitude of this excess.

These results were confirmed by Yang (1952) using perturbation theory, however, in this paper the main concern was the correct order of various limiting processes occurring. At the critical parameter value, the two largest eigenvalues of Onsager's V merge, and a true mathematical discontinuity only occurs for an infinite lattice. Yang & Lee (1952) also showed that all zeroes of the NC lie on the unit circle in the complex plane, conditioning on the value of some thermodynamic quantities.

Several different approaches to the analysis of Ising's model have not resulted in deriving new results since they are all variations of a general field theory approach. These are all based on Wick's Theorem which relates the trace of a product of linear sums or operators known as Clifford or Fermi operators to a mathematical object called a Pfaffian, a triangular array of numbers somewhat like a determinant, which almost has a cyclic structure leading to a closed form expression.

One property of the Pfaffian is that it can be expanded so that suffixes are grouped together in all possible pairs, but every suffix appears only once (in contrast to determinants). This leads to a matching problem of enumerating all ways of connecting lattice sites into non-overlapping pairs, which can be solved for all planar lattices (Kastelyn 1963). Other methods replacing Fermi with Bose operators or the Kac-Ward Determinant (Kac & Ward 1952) did not yield computationally advantageous routes.

Temperley (1972) catalogues properties of the two-dimensional Ising model that can be rigorously stated.

Relationship to the Gaussian model

Exact solutions are available for the Gaussian and spherical models in any dimension (Baxter 1982, Baker & Kawashima 1996). Here the Gaussian model is a reparameterization of the auto-Gaussian models of Besag (1974) introduced in Section 4.2.7 in equations (4.43)–(4.45). Baker & Kawashima (1996) notes that these are the only non-trivial statistical mechanical models where exact solutions are available in dimensions higher than two. I give a summary of their notes on the Gaussian model.

The Gaussian model originated (Kac 1964), as an attempt to simplify the Ising model in order to obtain a solution in any dimension for the NC and other thermodynamic properties. The binary random variables y_i are replaced with continuous-valued y_r defined on locations

$r = (r_1, r_2, \ldots, r_d)$ of a regular d-dimensional hypercubic lattice with L^d sites and with Hamiltonian

$$-k_B T \ln h(y) = -\sum_{r \neq s} J_{rs} y_r y_s - H \sum_r y_r. \tag{4.103}$$

Summations extend over all lattice sites. Assuming each y_r is distributed as a Gaussian with zero mean and variance σ_r^2, then the probability density is

$$p(y \mid k_B, T, J, H) = \frac{1}{c(\theta)} \exp\left\{ \frac{1}{k_B T} \sum J_{rs} y_r y_s + \frac{H}{k_B T} \sum y_r - \sum \frac{y_r^2}{\sigma_r^2} \right\} \tag{4.104}$$

and so the NC is

$$c(k_B, T, J, H) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(y \mid k_B, T, J, H)\, dy_r. \tag{4.105}$$

The simplest assumptions for simplifying the model comprise stationarity of the distribution over the lattice and periodic boundary conditions. Thus

$$J_{rs} = J(r - s) \qquad J(r + n) = J(r) = J(-r) \tag{4.106}$$

where $n = (n_1, n_2, \ldots, n_d)$ is the vector of lattice size over all dimensions. Exact evaluation of the NC proceeds by noting that the interaction parameters form a cyclic matrix (generalized), which can then be diagonalized using a Fourier transform. The exponent of the integrand above can be diagonalized using an orthogonal change of variables to yield a Gaussian integral in d dimensions. Furthermore since all $\{y_r\}$ are independent, the integral factors into separable components. However the free energy is not defined for temperatures smaller than a critical value which depends on σ^2: $T_c = dJ\sigma^2/k_B$.

The spherical model was then devised by Kac to overcome this difficulty as well as to incorporate more of the features of the Ising model. The constraint

$$\sum_r y_r^2 = L \tag{4.107}$$

was imposed. This is automatically satisfied by the Ising parameterization since $y_i = \pm 1$. The difference is that the spherical model allows uniform distribution of these spins on the sphere, whereas the Ising is restricted to the 2^L vertices of the hyperlattice.

Hyperscaling hypothesis

Researchers have also investigated whether the phase transition property or spontaneous magnetization occurs for higher dimensions. A large body of work in statistical physics has been concerned with proving the hyperscaling hypothesis for various models, including Ising models (See (Baker & Kawashima 1996) for an overview.) The Lebowitz 4-point inequality lies at the heart of the hypothesis relating the fourth and second order moments of site values: (Lebowitz 1974).

$$\mathrm{E}[y_i y_j y_k y_l] - \mathrm{E}[y_i y_j]\mathrm{E}[y_k y_l] \leq \mathrm{E}[y_i y_k]\mathrm{E}[y_j y_l] + \mathrm{E}[y_i y_l]\mathrm{E}[y_j y_k] \tag{4.108}$$

The two-dimensional Ising model is interesting because it satisfies the hypothesis, as proven in a series of papers (Stephenson 1964, Aizenman 1981).

It is interesting that the Gaussian model does not satisfy the hyperscaling hypothesis since there is no divergence of the Lebowitz 4-point inequality. This explains why it is the Ising or autologistic models instead of the Gaussian model, which have been applied to those situations where there is a clear change in behaviour for the spatial process beyond a certain degree of interaction between sites.

4.4 Simulation from MRFs

In this section we detail simulation methods for generating samples from binary Markov random field models, such as the Autologistic. Simulation underpins the Bayesian approach to inference used for the hierarchical models presented in Chapters 5, 6 and 7.

Independent samples have been difficult to obtain (Binder & Heermann 1997, for example) so we consider dependent sampling schemes. Dependent samples from the Autologistic model may be obtained using several variations of Markov Chain Monte Carlo (MCMC). Originating from seminal work by Metropolis et al. (1953) for the simple Ising model, MCMC has become popular when analytical or numerical techniques have been unsuccessful in exploring the probability distribution of interest. They harness tools from Markov Chain theory: given an equilibrium distribution π, a process can be constructed which converges in distribution to π, thus permitting simulation from π. The general MCMC computational and inference framework is outlined and defined in Section 4.4.1.

The two most popular approaches to construction are Gibbs sampling (Geman & Geman 1984) and Metropolis-Hastings (Hastings 1970). An introduction to details relevant to their implementation in Chapters 5–7 is provided in Sections 4.4.2 and 4.4.3 respectively. An implementation issue underlying the Metropolis-Hastings method is the choice of proposal distribution (Sections 4.4.2 to 4.4.2). An overview of other samplers is provided in Section 4.4.5 indicating future areas of exploration outside the scope of this thesis.

Where there is more than one parameter, or the parameter has more than one component, different samplers may be combined to simulate from a multi-dimensional parameter. Broad sampling strategies for combining samplers are outlined in Section 4.4.4 and later employed in Chapters 5–7.

Diagnostics reviewed and outlined in Section 4.4.7 are employed in later chapters to quantify the amount of dependence between simulations, and to establish whether equilibrium has been achieved to ensure simulations are representative of the distribution of interest. Other features of construction impact on the rapidity of convergence to π and dependence of simulations from π once equilibrium has been established. These are briefly discussed in Section 4.4.6.

The reader is referred to the many reviews of MCMC now available, such as Sokal (1989), Tierney (1991), Smith & Roberts (1993), Besag & Green (1993) and Gilks, Richardson & Spiegelhalter (1996).

4.4.1 Definition of Markov chain Monte Carlo

The aim of MCMC methods is to construct a Markov Chain which tends in distribution to the desired target density $\pi(x)$. For example, in a Bayesian context, $\pi(x)$ often corresponds to a posterior distribution $\pi(x \mid y)$.

We require a Markov Chain defined on parameter space $\mathcal{X}$ with transition probability density function $P(x, x^*)$ where

$$P^t(x, x^*) \xrightarrow{\mathcal{D}} \pi(x^*) \quad \text{as} \quad t \to \infty. \tag{4.109}$$

Here the parameter x may be one-dimensional or multi-dimensional, and may consist of several components, $x = (x_1, x_2, \ldots, x_k)$.

To ensure that the Markov chain produced in this way converges, two properties are required of the transition matrix: irreducibility and aperiodicity (Tierney 1991). A Markov chain defined by π is irreducible if every state is accessible at all times during the simulation.

That is, for every set A with $\pi(A) > 0$, the probability that the chain will enter state A is positive, regardless of the initial chain value $x^{(1)}$. A Markov chain is aperiodic if it does not cycle through a subset of states in with positive probability. There are a number of sufficient conditions that can be tested to ensure that these two conditions are met which depend on the sampler chosen.

In addition to satisfying these two requirements, an ideal transition function would be simple and converge *quickly* to the equilibrium distribution. Two methods of constructing these transition functions are Gibbs Sampling and various versions of the Metropolis, which are all variants of the more general Hastings algorithm (Metropolis et al. 1953, Hastings 1970, Geman & Geman 1984).

Once the chain has been constructed, we observe values of the parameter vector at each iteration: $\{x^{(t)} : t = 1, 2, \ldots, T\}$. An initial transient period, where the chain has not yet attained equilibrium, is eliminated before estimation proceeds. An estimator of a function f operating on the variable x, such as the mean or variance, may be evaluated:

$$\mathrm{E}_\pi[f] = \sum_{x \in \mathcal{X}} f(x)\pi(x). \tag{4.110}$$

The most straightforward estimator of $\mathrm{E}_\pi f$ is then the sample mean

$$\bar{f}_T = \frac{1}{T}\sum_{t=1}^{T} f(x^{(t)}) \tag{4.111}$$

By the ergodic theorem, and due to the irreducibility and aperiodicity of the sampler, this estimator converges almost surely to $\mathrm{E}_\pi f$ as $t \to \infty$ regardless of the starting value for the simulation $x^{(0)}$. See Besag & Green (1993) for a more rigorous treatment.

The rate of convergence, however, may depend on the starting value. Gelman & Rubin (1992) investigate the effect of different starting values on convergence, and devise a test statistic for combining chains beginning with different starting values $x^{(1)}$, to determine a suitable "burnin" time.

Two important steps in implementing an MCMC algorithm for conducting inference are initialization of the chain and iteration to produce the chain of simulated parameter values. Initialization requires initialization of each parameter component and selection of sampling mechanisms and associated distributions. Parameters on which the whole analysis is to be conditioned are identified and assigned fixed values. Parameters to be explored and estimated using MCMC are identified, and initialized to starting values $x^{(0)}$ corresponding to simulation time 0. An appropriate MCMC sampler is chosen to update each of these parameters within an overall sampling régime. A random sampling or systematic sampling régime are popular choices for describing how parameter components are updated at each iteration. Finally, the main body of the MCMC algorithm iteratively produces simulations for each of the parameters.

4.4.2 Metropolis-Hastings

The Metropolis-Hastings method refers to a general form described by Hastings (1970), based on the simplest version proposed by Metropolis et al. (1953). This method is based on a symmetric proposal distribution which eliminates some computational effort. Its application to the Ising model is discussed in more detail in Ripley (1988) and Hammersley & Handscomb (1964). The algorithm is an iterative scheme which involves 'proposing' new parameter values and accepting or rejecting them at each iteration.

The Metropolis-Hastings technique is most useful when it is easy to evaluate likelihood ratios of the distribution of interest $\pi(x) / \pi(x^*)$. It relies on good choice of an independently selected proposal distribution for which it is also easy to evaluate the likelihood ratios. This is exactly the situation that arises with the autologistic distribution, since the troublesome normalization constant can be omitted from these simulations. So, due to its simplicity and flexibility, the Metropolis-Hastings method is selected for providing simulations underlying inference for the autologistic distribution in this thesis.

At each iteration, the old value of a parameter X is used to generate a new proposed value x^* from a proposal distribution $r(x^* \mid x)$. The new proposed value x^* is accepted with probability $\alpha(x^* \mid x)$ given by

$$\alpha(x^* \mid x) = \min\left(1, \frac{\pi(x^*)r(x \mid x^*)}{\pi(x)r(x^* \mid x)}\right) \tag{4.112}$$

The value x^* is rejected and the original value of x retained with probability $1 - \alpha(x^* \mid x)$. This can be achieved by generating a uniform random variate on the unit interval

$$U \sim \text{Unif}[0, 1]$$

and accepting the new x^* if $U < \alpha(x^* \mid x)$, otherwise rejecting the new proposed value x^* and retaining the original value x.

Note that in the case of uniform or symmetric r the r-ratio above simplifies to unity. This is exploited in many practical applications of the Metropolis-Hastings sampler.

The choice of this proposal distribution r is not connected to π in general, so can be *model-independent.* Its associated variance can, however, affect the convergence rate of the MCMC. For practical reasons, the proposal distribution r should be chosen so that it is relatively easy to generate random variates from r and also relatively easy to evaluate likelihood ratios $\frac{r(x^* \mid x)}{r(x \mid x^*)}$. Reversibility and positivity of transition probabilities is required to ensure irreducibility and aperiodicity of the MCMC chain as discussed in the previous section.

The proposal distribution should also encourage sufficient exploration of $p(x)$. Visual inspection of the MCMC simulation can diagnose this to some extent, as well as ensuring acceptance ratios are in the range 0.3 – 0.7 (Best, Cowles & Vines 1995). It is necessary to balance a thorough examination of a portion of the sample space with wide coverage of the sample space.

Proposals for $[0, 1]$ data

Let us consider the case where the parameters to be estimated are bounded. For instance, in the *Dingo* case study, the parameters are probabilities which are defined on $\Omega_0 = [0, 1]$. Generally the best results are achieved when the proposal distribution chosen is symmetric and centred around the old parameter value. For bounded parameters however, if the old parameter value lies too close to the endpoints of the interval, then the new proposed value may be generated to lie outside this interval. In the following section we present several slight adjustments to symmetric proposal distributions suitable for $[0, 1]$ variates to ensure that the correct boundaries and properties are maintained.

Rotated uniform distribution

The uniform distribution can be altered to ensure that generated values, symmetrically distributed about the old value of the parameter, still lie within the unit interval. We wrap

the endpoints of the interval into a circular interval, with 1 being followed by 0.

To obtain a rotated uniform variate we first sample the intermediate value of the parameter, x^{**}, from a uniform distribution with half-width h, centred at the old parameter value, x

$$x^{**} \sim \text{Uniform}\,(x-h, x+h)\,.$$

Thus x^{**} may be obtained from a standard uniform variate $U \sim \text{Uniform}\,(0,1)$ via

$$x^{**} = 2hU + x - h$$

and since $\mathrm{E}[U] = \frac{1}{2}$ and $\mathrm{Var}[U] = \frac{1}{12}$, we have

$$\begin{aligned} \mathrm{E}[x^{**}] &= 2h\mathrm{E}[U] + x - h \\ &= x \\ \mathrm{Var}[x^{**}] &= 4h^2\mathrm{Var}[U] \\ &= h^2/3 \end{aligned} \tag{4.113}$$

If the intermediate value x^{**} lies outside the unit interval, we then *rotate* this value back into the interval to obtain the new parameter value x^{**}. That is,

$$x^{**} = \begin{cases} x^{**}+1, & \text{if } x^{**} < 0 \quad \textit{i.e. } U < \frac{h-x}{2h} \\ x^{**}\quad , & \text{if } 0 \le x^{**} \le 1 \quad \text{otherwise} \\ x^{**}-1, & \text{if } x^{**} > 1 \quad \textit{i.e. } U > \frac{3h-x}{2h} \end{cases}. \tag{4.114}$$

By symmetry, the probability density of generating the new from the old parameter value via the proposal distribution is

$$r(x^* \mid x) = \frac{1}{2h}. \tag{4.115}$$

This holds over the whole range of x since

$$\begin{array}{lll} x^* \in [0, x+h-1] \cup [x-h, 1] & \text{for} & 0 \le x-h < 1 \le x+h \\ x^* \in [0, x+h] \cup [1-(h-x), 1] & \text{for} & x-h \le 0 < x+h \le 1 \\ x^* \in [x-h, x+h] & \text{for} & 0 \le x-h < x+h \le 1 \end{array} \tag{4.116}$$

We also need to assume that $0 < h \le \frac{1}{2}$ for $r(x^* \mid x)$ to be a proper distribution function.

Therefore, the proposal ratio is always 1 since

$$\frac{r(x \mid x^*)}{r(x^* \mid x)} = \frac{1}{2h} \Big/ \frac{1}{2h} = 1 \tag{4.117}$$

However, one disadvantage of the rotated uniform as a proposal distribution is that it can be somewhat inefficient to sample new values of the parameter far removed from the original value, as occurs when an intermediate value outside the unit interval is obtained. The advantage of this is the resulting potential for increased movement around the proposal space.

Truncated uniform distribution

The truncated uniform overcomes the inefficiency of the rotated uniform in sampling far from the current parameter value but loses its symmetry. The proposal ratio is no longer unity for all pairs of old and new parameter values.

To obtain a truncated uniform variate, we sample from a uniform distribution with half-width h, centred at the old parameter value x, truncated to the unit interval

$$x^* \sim \text{Uniform}\left(\max(0, x-h), \min(x+h, 1)\right). \tag{4.118}$$

The width of this uniform distribution is

$$\begin{array}{ll} \frac{1}{h+x}, & \text{for } x-h \leq 0 < x+h \leq 1 \\ \frac{1}{h+h}, & \text{for } 0 \leq x-h \leq x+h \leq 1 \\ \frac{1}{h+(1-x)}, & \text{for } 0 \leq x-h < 1 \leq x+h \end{array} \tag{4.119}$$

which may simplified to

$$h + \min(h, x, 1-x) \quad 0 \leq x^* \leq 1. \tag{4.120}$$

Hence a new proposed value for the parameter may be generated from a standard uniform variate $U \sim \text{Uniform}\,(0,1)$ via

$$x^* = (h + \min(h, x, 1-x))\, U + \max(0, x-h) \tag{4.121}$$

Thus the probability density of the transition from the old to the new parameter value is simply

$$r(x^* \mid x) = \frac{1}{h + \min(h, x, 1-x)} \quad 0 \leq x^* \leq 1 \tag{4.122}$$

The proposal ratio is therefore

$$\frac{r(x \mid x^*)}{r(x^* \mid x)} = \frac{h + \min(h, x, 1-x)}{h + \min(h, x^*, 1-x^*)} \tag{4.123}$$

which only becomes unity when $x^* \in \{x, 1-x\}$.

Normal distribution

Although this distribution is not a suitable choice as a proposal distribution for $[0,1]$ data, it introduces discussion of the truncated Normal distribution which is suitable. The truncated Normal is also useful as the proposal distribution for logit transformed $[0,1]$ data. The uniform distributions already considered give equal weighting to proposed values of the parameter which lie within a specified width of the old value. The advantage of the Normal as a proposal distribution is that it places more weight on proposed values which lie *closer* to the old value.

The new proposed value of the parameter is sampled from a Normal distribution centred at the old value of the parameter with specified variance σ^2

$$x^{**} \sim \text{N}\left(x, \sigma^2\right). \tag{4.124}$$

We may generate x^{**} from a standard normal variate $U \sim \text{N}\,(0,1)$ with cumulative density function $\Phi(u)$ via

$$x^{**} = \sigma U + x. \tag{4.125}$$

The normal proposal density is

$$r(x^* \mid x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\left(\frac{x^* - x}{2\sigma} \right)^2 \right\} \tag{4.126}$$

where $-\infty \leq x^* \leq \infty$. Now the proposal ratio used in the Metropolis-Hastings algorithm for the acceptance probability is

$$\frac{r(x \mid x^*)}{r(x^* \mid x)} = 1 \tag{4.127}$$

since the squared terms cancel, and σ is constant.

Truncated normal distribution

Both the rotated and the truncated uniform distribution give equal weighting to all proposed parameter values which lie within a specified width of the old value. The normal proposal distribution places more weight on proposed values which lie *closer* to the old value, and can be truncated to the unit interval in this case.

The new proposed value of the parameter is sampled from a Normal distribution centred at the old value of the parameter with a specified variance σ^2

$$x^{**} \sim \mathrm{N}\left(x, \sigma^2\right) \tag{4.128}$$

We may generate x^{**} from a standard normal variate $U \sim \mathrm{N}(0,1)$ with cumulative density function $\Phi(u)$ via

$$x^{**} = \sigma U + x \tag{4.129}$$

If the generated value of x^{**} lies outside the unit interval, then follow a rejection sampling approach and discard it and resample. Thus

$$\Pr\left(x^{**} < 0\right) = \Pr\left(x^{**} > 1\right) = 0 \tag{4.130}$$

Truncating the density of x^{**} to the unit interval requires a slight adjustment to maintain $r(.)$ as a proper probability density function, giving the truncated normal version of the proposal density as

$$r(x^* \mid x) = \frac{1}{\left\{\Phi\left(\frac{1-x}{\sigma}\right) - \Phi\left(\frac{-x}{\sigma}\right)\right\}\sqrt{2\pi}\sigma} \exp\left\{-\left(\frac{x^* - x}{2\sigma}\right)^2\right\} \tag{4.131}$$

where $0 \leq x^* \leq 1$. Now the proposal ratio used in the Metropolis-Hastings algorithm for the acceptance probability is

$$\frac{r(x \mid x^*)}{r(x^* \mid x)} = \frac{\Phi\left(\frac{1-x}{\sigma}\right) - \Phi\left(\frac{-x}{\sigma}\right)}{\Phi\left(\frac{1-x^*}{\sigma}\right) - \Phi\left(\frac{-x^*}{\sigma}\right)}. \tag{4.132}$$

4.4.3 Gibbs Sampling

In Gibbs sampling it is the conditional distributions of each component of the parameter vector which are required. This can be particularly useful for problems which are more easily expressed in terms of these conditional distributions.

The Gibbs sampler proceeds by updating the parameter by randomly sampling from its conditional distribution keeping all other parameters constant (as denoted by ...)

$$x \sim \pi(x| \ldots). \tag{4.133}$$

The Gibbs sampler is actually a special case of the Metropolis-Hastings algorithm where the acceptance step is avoided, since the proposal distribution is chosen so that the proposal ratio cancels out of the posterior density ratio

$$r(x^*|x) = \pi(x|\ldots) \tag{4.134}$$

and so the acceptance probability becomes one. The cost for this is that the new proposed value of the parameter must be sampled from the posterior density of the parameter, which can be computationally expensive unless this density is a standard distribution, has a conjugate prior, or only takes a few values. Thus in contrast to the Metropolis-Hastings algorithm with a symmetric proposal distribution, this is a *model-dependent* choice of proposal distribution.

4.4.4 Hybrid strategies

When the parameter has dimension greater than one and comprises several components various strategies can be applied for simulating from the ensemble of components. The sampling strategy may vary and be tailored for each individual component, giving rise to a hybrid sampling strategy (Tierney 1994). Hybrid strategies are used during implementation of the MCMC computational approach to Bayesian inference used from Chapter 5 onwards.

Hybrid strategies most often fall into two categories: mixtures and cycles. Under a mixture strategy, there are positive probabilities $\alpha_1, \alpha_2, \ldots, \alpha_m$ of using Markov transition kernels $P_1, P_2, \ldots, P_m$ with the same equilibrium distribution. In a cycling or systematic strategy, each of these transition kernels is used in turn.

The Gibbs sampler is often applied to multi-dimensional parameters via application of a simple cycling strategy. In an iteration, each parameter component is updated based on the most up-to-date values of all other components. So if the value of the parameter vector at the $t-1$st iteration is $x^{(t-1)} = (x_1^{(t-1)}, x_2^{(t-1)}, \ldots, x_K^{(t-1)})$, then in the next iteration of the algorithm, iteration t, the components are generated from the following conditional distributions:

$$\begin{aligned}
x_1^{(t)} &\sim \pi(x_1 \,|\, x_2^{(t-1)}, \ldots, x_K^{(t-1)}) \\
x_2^{(t)} &\sim \pi(x_2 \,|\, x_1^{(t)}, x_3^{(t-1)}, \ldots, x_K^{(t-1)}) \\
&\vdots \quad\quad \vdots \\
x_k^{(t)} &\sim \pi(x_k \,|\, x_1^{(t)}, \ldots, x_{k-1}^{(t)}, x_{k+1}^{(t-1)}, \ldots, x_K^{(t-1)}) \\
&\vdots \quad\quad \vdots \\
x_K^{(t)} &\sim \pi(x_K \,|\, x_1^{(t)}, \ldots, x_{K-1}^{(t)})
\end{aligned}$$

Another common use of a hybrid strategy is to combine Gibbs with occasional steps from a Metropolis-Hastings chain in a mixture or cycle. This essentially 'restarts' the Gibbs sampler, reducing correlations and the tendency to get 'stuck' at local maxima.

An approach used by Zeger & Karim (1991) partitions a multidimensional parameter $x = (x_1, x_2)$. This is an advantage when it is easy to directly sample from the conditional distribution of $x_1 \,|\, x_2$, but not for $x_2 \,|\, x_1$. Then a Gibbs sampler may be used for updating $x_1 \,|\, x_2$, and a Metropolis sampler may be used for updating $x_2 \,|\, x_1$.

Theoretical results show that hybrid strategies lose none of the properties of the basic strategy (Besag & Green 1993). Tierney (1991) notes that if one sampler in a mixture is

irreducible and aperiodic, then so is the mixture sampler. Conversely, if one sampler in a cycle is irreducible and aperiodic, then *often but not always* the cycle sampler is also.

Extensions to these hybrid strategies mentioned in Smith & Roberts (1993), Gelfand & Smith (1990) include updating groups of parameter components in one step, thus drawing from a multivariate conditional distribution. Convergence could be faster if several components are highly correlated and it is not too costly to sample from the multivariate conditional density. Components can be updated in random order rather than systematic order.

4.4.5 Other samplers

Samplers other than the Gibbs and Metropolis-Hastings samplers are presented here to show the wide range of choices available and to compare features with these samplers.

Recent developments in algorithms for simulation of Markov random fields for physical modelling, such as the Ising model, are reviewed in Binder & Heermann (1997). This text gives a thorough coverage of important results derived in the statistical physics context, although its scope is wider than required for this thesis. There are also some new results in the statistical literature. I give a brief overview of these results as they relate to simulation of the Ising/autologistic distributions. Simulation methods include: Gibbs (site flipping), block updates, cluster algorithms, simultaneous estimation at various scales, reweighting via single histogram or multiple histogram methods, simulated tempering, perfect simulation and vectorization.

In a statistical physics context, Hammersley & Mazzarino (1983) propose a sampling regime for the simple Ising model based on a Gibbs sampling update, which they call *spin flipping*. Of interest is the number of iterations in their MCMC chain: on a 100×100 square lattice, they use a burnin of 1,000,000 flips and a simulation time of 25,000,000 flips. This contrasts with choices made by other authors from other disciplines, where as little as 200 iterations are used (Geyer 1994).

A variation is to exchange neighbours, rather than to flip individual sites. The *nearest neighbour exchange algorithm* (Kawasaki 1972) selects the next pair of sites to update, then proposes whether to exchange their values or not.

Block updates can improve the rate of convergence by increasing the rate when there is little variability in the likelihood ratio at each MCMC iteration. Binder & Heermann (1988) outline block updating in more detail. These blocks are defined by their location on the lattice, irrespective of site values.

Cluster algorithms (Swendson & Wang 1987) update clusters of sites rather than single sites. Clusters are defined not by their location on the lattice, but by grouping together similar site values. These algorithms reduce the problem of critical slowing down during simulation, which leads to divergence in time taken to reach the equilibrium distribution[13]. As this time increases, so too does the statistical variation in estimates of functions. Heuristically this can be explained as follows. As the spatial interaction parameter increases, so too does the long range correlations between sites. Due to presence of large clusters of correlated sites, it takes a long time to "disintegrate" these clusters using single site updates. Use of cluster updates leads to a faster change in cluster pattern. The challenge is in the construction of appropriate clusters.

The *cluster flipping algorithm* (Glauber 1963, Swendson & Wang 1987) views the lattice as containing clusters of presence and absence. At each iteration a cluster is selected, and

[13]known as relaxation time in statistical physics (Binder & Heermann 1997, Geman & Geman 1984).

a proposal made whether to 'flip' the value from presence to absence or vice versa. These two methods assume evenly distributed lattices with $\mathrm{E}[X] = 0.5$, although they can be embedded into an algorithm containing another layer to update the number of presences V_0.

Swendson & Wang (1987) changes focus from geometrical clusters of presence/absence and allows the external field H to vary from positive to negative values, equivalent to allowing the prevalence to vary from low to high values. Physical clusters are defined by active bonds. These active bonds cannot occur between sites having different values. Given a pair of sites with the same values, then the probability of an active bond between them is

$$p(\text{active bond}) = 1 - \exp\left\{-\frac{2J}{k_B\mathcal{T}}\right\} = 1 - \exp\left\{-\frac{\theta_1}{2}\right\}. \tag{4.135}$$

Then the Normalization constant can be expressed as follows

$$\begin{aligned} c(\theta) &= \sum_{y} \exp\{-\beta\mathcal{H}\} \\ &= \sum_{\substack{\text{active bond} \\ \text{configurations}}} p^{N_b}(-1)^{N_m}2^{N_c} \end{aligned} \tag{4.136}$$

where N_b is the number of active bonds on the lattice, N_m is the number of missing bonds on the lattice, N_c is the number of clusters formed by these bonds. A similar expression was derived for the Potts model.

Within a Metropolis-Hastings style of sampler, the simulation algorithm proceeds for each iteration as follows. First active bonds are attached between pairs of sites with the same value according to acceptance probability defined by

$$U \leq p(\text{active bond between nearest neighbours})$$

where $U \sim \text{Unif}[0, 1]$; otherwise a missing bond is assigned. Clusters comprising homogeneous presence or absence are formed by these bonds. At random, a new value of presence or absence is assigned to clusters. This design is ergodic and satisfies the detailed balance equations. However, this assumes that the number of presences/absences remains constant.

Wolff (1989) devised a *single cluster* variation of the cluster algorithm (for the simple Ising model) which instead of first constructing all bonds on the lattice and then selecting a cluster, a single site is selected and the associated cluster constructed. This method reduces the time to convergence to the equilibrium distribution. In each iteration a site is selected at random, and bonds are created successively to its neighbours using the active bond probability given above. This continues until the cluster produced reaches its limit (no more active bonds for any outer sites). A software program to implement this algorithm is provided in Wang & Swendsen (1990).

Methods employing *simultaneous estimation at various scales* (as described in (Binder & Heermann 1997)) have a large drawback in computational efficiency for estimating the quantities of interest. In general these are the first and second derivatives of the log NC which correspond to the marginal mean and the Hessian matrix of the NC.

Reweighting (Binder & Heermann 1992) is a method which takes advantage of the fact that simulation of one particular system state z gives information on other nearby states. The simplest reweighting scheme is called the *single histogram method* and focuses on the

distribution of the energy function $V(z)$ for different values of parameters (temperature). This amounts to importance sampling the NC of the joint distribution of $\mathcal{H}$ and $\mathcal{T}$ using importance sampling function equivalent to the NC corresponding to temperature $\mathcal{T}_l$.

$$p(\mathcal{H},\beta) = \frac{p(\mathcal{H},\mathcal{T}_l)\exp\left\{-\left(\frac{1}{\mathcal{T}}-\frac{1}{\mathcal{T}_l}\right)\frac{\mathcal{H}}{k_B}\right\}}{\sum_{\mathcal{H}} p(\mathcal{H},\mathcal{T}_l)\exp\left\{-\left(\frac{1}{\mathcal{T}}-\frac{1}{\mathcal{T}_l}\right)\frac{\mathcal{H}}{k_B}\right\}} \tag{4.137}$$

This expression requires that the support of $p(\mathcal{H}|\mathcal{T}_l)$ to be sufficiently broad permitting exploration of the parameter space. The tails of this distribution can be badly estimated so another approach is to combine several histograms at suitably chosen temperatures (or other control parameters). This is called the *multiple histogram extrapolation* method (Binder & Heermann 1992), and improves accuracy. Some recent advances (Binder & Heermann 1997) have employed this histogram technique for when the marginal mean is not 0.5 (non-zero external field). Reweighting within simulation updates was called *umbrella sampling* by Torrie & Valleau (1977).

Simulated tempering (Geman & Geman 1984) 'expands the ensemble', which in statistical terms, means introducing latent variables (Dempster et al. 1977). This also corresponds to the reversible jump MCMC approaches of Green (1995). Transition rules need to be specified for jumping from one temperature to another. The *broad histogram method* (Binder & Heermann 1997) focuses on sampling the Gibbs distribution directly.

The *stochastic relaxation method* outlined in Winkler (1995), derived from the method of Geman & Geman (1984), is a form of simulated annealing and focuses on changing the time parameter $\mathcal{T}$ from equation (4.57). This approach augments the variable space by expressing β in its thermodynamic parameterization $1/(k\mathcal{T})$. Ripley's clock method (Binder 1986, Müller-Krumbhaar & Binder 1973, Ripley 1988) uses regeneration ideas (Mykland, Tierney & Yu 1992) in order to determine the next time that a large-scale change in the lattice should occur, with small scale changes (site flipping) occurring at in-between times.

Vectorization is another method of improving simulation. This is a computational technique which permits parallel calculations. The challenge is to identify which calculations are independent at any point during iterations, in order to select which calculations may occur in parallel. An in-depth treatment is given in Binder & Heermann (1992).

One way to ensure that optimal use is made of simulated values is to accurately identify when convergence to equilibrium has occurred. *Simulation time before convergence* to the equilibrium distribution (relaxation time) has been estimated for the simple Ising and some Potts models by Landau & Tang (1988).

Another way to identify that equilibrium has been reached is the *Perfect sampling technique*. Green & Murdoch (1999) refers to new work (Häggström & Nelander 1997*a*, Häggström & Nelander 1997*b*, Häggström, van Lieshout & Møller 1999) which explores application of the Perfect sampling technique (Propp & Wilson 1996) to Markov random field simulation. This requires simulation of the past to the origin at time 0 when the desired equilibrium distribution 'was' achieved.

4.4.6 MCMC performance: standard error

Ideally we require a chain which firstly converges rapidly to the equilibrium, and secondly has high estimation precision. Green & Han (1990) show that unfortunately these two requirements are in direct conflict with each other. It is therefore necessary to strike the correct balance between rapid convergence and high estimation performance. One method

of balancing these diverse requirements is to switch between different transition mechanisms designed to satisfy each requirement separately. Green & Han (1991) suggest using a process chosen for rapid convergence to equilibrium for the first T_0 iterations, and then use using a second process chosen for better estimation performance for the remaining T updates.

MCMC standard error is of pivotal interest both for quantifying error of derived estimators and for determining run length. Simulations obtained using MCMC will necessarily be dependent. Thus the independent sample estimate of variance s^2 will not be accurate. Other methods which take into account the dependence between simulated values can be borrowed from Time Series methods. The simplest method takes into account only the first order autocorrelation ρ_1. The Integrated Autocorrelation Time method is more complex and takes into account all practically significant autocorrelations. We outline these briefly below.

The usual simulation standard deviation ignores dependence between samples:

$$\text{s.e.}_{\text{IND}}[f] = \sqrt{\frac{1}{T(T-1)}\sum_{t=1}^{T}(f^{(t)}(x)) - \overline{f})^2}. \tag{4.138}$$

where $\overline{f} = \sum_{t=1}^{T} f(x^{(t)})$ is the simulation average of the function f.

A simple AR(1) estimate (Tierney 1991) is

$$\text{s.e.}_{\text{AR(1)}}[f] = \text{s.e.}_{\text{IND}}\sqrt{\frac{\widehat{\rho}+1}{\widehat{\rho}-1}} \tag{4.139}$$

where $\widehat{\rho}$ is the usual estimate of the lag 1 autocorrelation of $\{f(x^{(t)})\}$.

Finally the Integrated Autocorrelation Time (IACT) estimate (Sokal 1989, Green & Han 1990) incorporates information on the most important autocorrelations:

$$\begin{aligned}
\text{Var}\left[\bar{f}_T\right] &= \frac{1}{T^2}\sum_{s=1}^{T}\sum_{t=1}^{T}\text{Cov}\left[f(x^{(s)}), f(x^{(t)})\right] && (4.140)\\
&\approx \frac{\sigma^2}{T}\sum_{t=-(T-1)}^{T-1}\{1-\frac{|t|}{T}\}\rho_t(f) && (4.141)\\
&\approx \frac{\sigma^2}{T}\tau(f) && (4.142)
\end{aligned}$$

So, expressing the MCMC standard error similarly to 4.139 we have

$$\text{s.e.}_{\text{IACT}}[f] = \text{s.e.}_{\text{IND}}\sqrt{\widehat{\tau}(f)}. \tag{4.143}$$

They call $\tau(f)$ the integrated autocorrelation time (IACT). In practice we must estimate the value of $\tau(f)$. Three different estimators are provided in the literature for estimating the integrated autocorrelation time. These are a naive estimator, a truncated periodogram estimator preferred by Green & Han (1990), and a spectral density estimator similar to Priestley (1981).

IACT: Naive estimator

The naive estimator of IACT is

$$\widehat{\tau}(f) = \sigma^2\sum_{t=-\infty}^{\infty}\widehat{\rho}_t(f) \tag{4.144}$$

where $\widehat{\rho}_t(f)$ is the sample autocorrelation function

$$\widehat{\rho}_t(f) = \frac{\widehat{c}_t(f)}{\widehat{c}_t(0)} \tag{4.145}$$

$$\widehat{c}_t(f) = \frac{1}{T-|t|}\sum_{i=1}^{T-|t|}(f_i - \bar{f})(f_{i+|t|} - \bar{f}) \tag{4.146}$$

However it can be shown (Sokal, 1989) that this estimator is badly biased, and inefficient in the sense that the variance does not decrease to zero as sample size increases. So more sophisticated estimators are required.

IACT: Truncated periodogram estimator

This estimator is advocated by Sokal (1989) and Green and Han (1991). It is based on a result (Priestley: 1981,p225) which relates IACT to the spectral representation of the process:

$$\tau(f) = \sigma^2 \sum_{t=-\infty}^{\infty} \rho_t(f) = 2\pi f(0) \tag{4.147}$$

where $f(0)$ is the spectral density function of the process at frequency 0. This gives the truncated periodogram estimator:

$$\widehat{\tau}(f) = \sum_{t \le M} \widehat{\rho}_t(f) \tag{4.148}$$

with window width M chosen adaptively to be the minimum integer such that

$$M \ge c\widehat{\tau}(f) \tag{4.149}$$

where c is some constant. Sokal suggests a value of $c = 2$ if the sample autocorrelation function is roughly a pure exponential, although if it decays more slowly, then values of c between 3 and 5 may be more appropriate. The proviso is that the sample size is over 1000 times the IACT. Green & Han (1991) find that a value of $c = 3$ was adequate for their application.

IACT: Spectral density estimator using Bartlett window

Hastings (1970) recommends a form of the spectral density estimator using the Bartlett window, as presented by Priestley (1981,p439). The series is divided into b separate consecutive blocks of length m. The between blocks mean square is

$$\text{BBMS} = \frac{m}{b-1}\sum_{i=1}^{b}\left(\left\{\frac{1}{m}\sum_{t=(i-1)m+1}^{im} f(x^{(t)})\right\} - \bar{f}_T\right)^2 \tag{4.150}$$

which is approximately an unbiased estimator of $\sigma^2\tau(f)$ as $b, m \to \infty$. The drawback to this approach is that the choice of the block size b affects the performance of the estimator.

Run length

An important design issue is to choose between a single long simulation run or several shorter runs, keeping in mind that equilibrium will not be attained in a run that is too short. The problems with using shorter runs are that: the shorter the run, the more difficult it is to determine whether the run is sufficiently long; and it is an inefficient use of the data, especially after accounting for the size of the initial transient which must be discarded. The disadvantages of using longer runs are that variances of estimates are more difficult to obtain, due to dependence between observations.

Estimates of MCMC standard error can be used to estimate run length required to ensure equilibrium has been achieved in the chains. These include an aggregation in pairs method due to Flyvbjerg & Petersen (1989) and methods involving reweighting the sample size by the MCMC standard error (Tierney 1991, Green & Han 1990).

Binder & Heermann (1997) reviews a method due to Flyvbjerg & Petersen (1989) which endeavours to estimate how functions to be estimated can be aggregated to produce effectively independent values for estimation. Suppose we have an *iid* sample of size T from equilibrium distribution π, then the standard error of the sample mean of a function of the parameters $f(x)$ is $\sigma(f)/\sqrt{T}$, where $\sigma(f)$ is the posterior standard deviation of $f(x)$. With technique, simulated values $f^{(1)}, f^{(2)}, \ldots, f^{(T)}$ can be aggregated into blocks via

$$f_*^{(i)} = \frac{1}{2}\left(f^{(2i-1)} + f^{(2i)}\right). \tag{4.151}$$

Then the number of samples in f_* are halved $T_* = T/2$ but the sample means are equivalent $\overline{f} = \overline{f_*}$ as are the variances $\sigma^2(f) = \sigma^2(f_*)$. However the covariances $\gamma_{ij} = \mathrm{Cov}\left[f^{(i)}, f^{(j)}\right]$ between observations $f^{(i)}$ and $f^{(j)}$ are reduced. Assuming stationarity of covariances along the chain with $\Delta t = |i - j|$

$$\gamma_*(i,j) = \gamma_*(\Delta t) = \begin{cases} \frac{1}{2\gamma(0)} + \frac{1}{2\gamma(1)} & \Delta t = 0 \\ \frac{1}{4\gamma(2\Delta t - 1)} + \frac{1}{2\gamma(2\Delta t)} + \frac{1}{4\gamma(2\Delta t + 1)} & \Delta t > 0 \end{cases} \tag{4.152}$$

The variance for f, f_*, *etc*, can be estimated by

$$\sigma^2(f) = \frac{1}{T^2}\sum_{i,j=1}^{T}\gamma(i,j) = \frac{1}{T}\left[\gamma(0) + 2\sum_{i=1}^{T-1}(1 - \frac{t}{T}\gamma(t)\right] \tag{4.153}$$

and so is bounded below by $\sigma^2(f) \geq \gamma(0)/T$. The authors suggest that the series $\sigma^2(f), \sigma^2(f^*), \sigma^2(f^{**}), \ldots$ is examined for a plateau. As the number of blockings applied increases, then the estimate $\sigma^2(f)$ approaches a constant finite value, which coincides with essentially independent blocked observations f.

Tierney (1991) suggests modelling our dependent series as a simple AR(1) process, so that the asymptotic standard error of the sample mean would be

$$\frac{\sigma(f)}{\sqrt{T}}\sqrt{\frac{1+\rho_1}{1-\rho_1}}. \tag{4.154}$$

A rough estimate of ρ_1 may then be used to adjust the sample size.

$$N^{\text{new}} = \frac{1+\rho_1}{1-\rho_1}N \tag{4.155}$$

Since we have $\frac{T}{\widehat{\tau}(f)}$ essentially independent observations, preliminary estimates of $\widehat{\tau}(f)$ may be used to adjust the sample size to the desired level. For instance, Aykroyd & Green (1991) suggest that the equivalent of 100 independent observations is sufficient in practice. Hence the general rule-of-thumb given by these authors for choosing the sample size is to ensure that

$$\widehat{\tau}(f) \leq 0.01T \tag{4.156}$$

which implies negligible contribution from MCMC variation to the variance of the estimator.

4.4.7 Diagnostics

We know that in theory (Besag & Green 1993) MCMC chains will converge to the desired posterior distribution of the parameters. In practice however we need general procedures to determine whether a particular realisation of a chain has indeed converged to the desired distribution.

MCMC diagnostics test or investigate whether simulations are from the equilibrium distribution of interest. They are therefore an essential step supporting inference of parameters in the hierarchical models and normalization constants of binary Markov random fields presented in this thesis.

This section briefly reviews the literature on practical application of MCMC techniques, and the problem of diagnosis of convergence to the equilibrium distribution of interest. Diagnostic methods range from visual inspection or "eyeball" plots of MCMC chains, monitoring indicator statistics obtained from the MCMC chains, to chain analysis via hypothesis testing. The reader is referred to standard texts on the topic for more details (Gilks et al. 1996, Sokal 1989, Cowles & Carlin 1996, Best et al. 1995, Mengersen, Roberts & Guihenneuc-Jouyaux 1999). Diagnostic tests are now sufficiently standardized that they are implemented in accessible software packages such as CODA (Best et al. 1995).

Before embarking on statistical tests, it is best to start with simple visualization and exploratory data analysis techniques.

- Visual inspection of plot of trace (time-series) of simulations will highlight 'Stickiness' or high autocorrelation in the chain and give some indication of modes visited.

- Several standard error estimates as discussed above in Section 4.4.6 give an indication of autocorrelation in the series. Similar to the IACT (Section 4.4.6), lag-1 autocorrelation (equations (4.154)–(4.155)) can assist in estimating the effective number of independent samples available for estimating the probability distribution of the values. In particular autocorrelation function plots help determine the stickiness of the chain. High autocorrelation at short lags would indicate that further thinning (and thus savings on computing space) is advisable; or alternatively that mixing is not sufficient and perhaps the design of the sampler should be adjusted, *e.g.* increase the variance or bandwidth of proposal distributions for Metropolis-Hastings samplers.

- Summary statistics of the Markov Chain produced may be monitored to ensure that convergence is being attained. Gelfand & Smith (1990) suggest following the 5%, 25%, 50%, 75% and 95% empirical quantiles of the posterior density over iterations. Several quantiles of the series would give a better overall picture than just following the mean.

- Smith & Roberts (1992) advocate investigation of the robustness of model choice by examining several perturbations of the model: omitting subsets of observations to determine influence on model; changes to the likelihood distribution; and changes to the prior distribution.

Some generic diagnostic tests have been incorporated into a software module called CODA (Best et al. 1995) for the statistical package Splus (Becker et al. 1988). These may be applied to diagnose convergence for output from any MCMC algorithm. These diagnostics include:

- Geweke (1992) first divided the data into two sections, by default the first 10% and last 50% of data. Then using a simple standard normal statistic, the means of each batch are compared. At the 5% significance level, if Geweke's statistic exceeds 1.96, then the null hypothesis that the batch means are equal should be rejected.

- Gelman & Rubin (1992) devised a test based on at least two independent chains each initialized with different starting values. The within chain and between chain variability is compared using ANOVA. Shrink factors are computed to represent the degree to which the chain's variance might be shrunk to attain the asymptotic variance. This method is very expensive in terms of computing time and space requirements, which are at least doubled.

- The test of Raftery & Lewis (1992) detects whether convergence has been achieved by assessing the quality of quantitle estimation. The optimum sample-size, to ensure a particular quantile is estimated to the desired degree of accuracy, can be computed. The default option is to determine how many samples are required in order to estimate the 2.5th percentile to within ± 0.005 of the "true" value with 95% confidence.

- Under the null hypothesis this statistic has a standard Normal distribution. The half width test compares the mean to the standard error, represented by the half-width of the 95% confidence interval. A CUSUM series (described below) may be standardised and used to construct a Brownian Bridge process (see Schruben 1982,1983). Hiedelberger & Welch (1983) base their test for stationarity on this Brownian bridge theory and use a Cramer-von-Mises statistic to test the null hypothesis that the data form a stationary process, with respect to the mean. Their half-width test examines the half-width of the 95% confidence interval for the mean, so is a test for excessive variability. They suggest successively deleting 0%, 10%, 20%, 30%, 40% and 50% of the series, calculating this statistic at each stage, and stopping when the remaining part of the series passes the test. If all tests fail, then a longer run is required, and the tests should be repeated. This strategy would be useful regardless of the statistical test being used. Larger values of the Cramer-von-Mises statistic indicate larger deviations from pure Brownian motion, and therefore an underlying non-stationarity.

Other useful diagnostics were not implemented in CODA, mostly since they post-dated the bibliographic review (Cowles & Carlin 1996). I found these were easily implemented in Splus (Becker et al. 1988), and use these to supplement the diagnostics listed above.

- With the Metropolis-Hastings algorithm, it is also important to monitor the acceptance rates, the proportion of times new values of the parameter generated from the proposal distribution are accepted. Acceptance rates can be computed over simulation

time and plotted as a time series to assist with assessing when the chain has reached equilibrium. According to Besag, Green, Higdon & Mengersen (1995) values within the range 30%–70% are adequate to ensure sufficient mixing (*i.e.* efficiency) of the sampler. Acceptance rates which are too low (*i.e.* less than 30% according to this rule of thumb) indicate that the algorithm is fairly inefficient. High acceptance rates (eg over 70%) indicate that the algorithm is fairly efficient. However it may also indicate that the algorithm does not move very quickly through the parameter space, possibly suggesting a slow convergence rate.

- The integrated correlation time (Green & Han 1990) can be computed to check that run length is sufficient. Its value can be used to estimate the number of effectively independent values available in the simulated chain. Green & Han (1990) suggest that the IACT should be small enough to ensure that the effective number of IID samples should be at least one hundred, *i.e.* $IACT \leq \frac{T}{100}$.

- CUSUM plots, suggested by Yu & Mykland (1994), can be used as an adjunct to the tests of Geweke and Hiedelberger & Welch, to ascertain whether the posterior means have drifted more than a typical "benchmark" chain might be expected to. They take the standard definition of a cumulative sum popular in the Quality literature and compare the time series of CUSUMs computed from the data to a CUSUM computed from an "ideal" simulated process having the same mean and variance as the data. The hypothesis being tested is that the means, μ_i, are all identically equal to some unknown parameter μ. The kth cumulative sum is

$$S'_k = \sum_{i=k+1}^{t} X^{(i)} - \mu \tag{4.157}$$

  Since μ is usually unknown, it may be estimated by the sample mean, and these cumulative sums may then be estimated by their empirical equivalent CUSUMs

$$\begin{aligned} \text{CUSUM}_k &= \sum_{i=k+1}^{t} X^{(i)} - (t-k)\bar{X} \\ &= -(t\bar{X} - \sum_{i=k+1}^{t} X^{(i)}) + k\bar{X} \\ &= -\sum_{i=1}^{k} X^{(i)} + k\bar{X} \\ &= -(S_k - k\bar{X}) \end{aligned} \tag{4.158}$$

  where $S_k = \sum_{i=1}^{k} X^{(i)}$ is the ith partial sum of the series $X^{(i)}$. An informal test is to compare the maximum drift of the benchmark to that of the data, and to compare the time periods over which these drifts occur. Sustained downward or upward movement of the CUSUMs away from the mean indicate non-stationarity, and should reflect results from applying the Hiedelberger & Welch tests and Geweke's test.

- The hairiness diagnostic (Brooks 1997) quantifies the opposite of the stickiness of the chain, and indicates the sampler's ability to traverse the parameter space. Together with the autocorrelation function plots, and to some extent the suite of CUSUM/Geweke/Hiedelb & Welch tests, this indicates whether the chain is mixing adequately and/or whether

thinning is necessary. Values below 0.5 in magnitude are believed to represent an adequate level of hairiness. Confidence intervals for the diagnostic were proposed by Brooks (1997). These confidence intervals need to be interpreted with care, since large sample sizes obtained as the time period gets longer, lead to deceptively tight confidence intervals.

Theoretical results (*e.g.* Meyn & Tweedie (1993)) are available for Markov chains which can assist in determining the time to convergence for particular models. This approach to determining relaxation time to equilibrium is model-dependent in contrast to the empirical approaches reviewed above. For highly complex systems, the empirical approaches may continue to be the only methods available for assessing convergence. Mengersen et al. (1999) review some of these more complex methods for diagnosing convergence which need to be tailored to the specific target distribution being simulated.

Discussion

It is important to note that some prior transformation of chains may be necessary to ensure the best performance of these diagnostics. For example, the logit or probit transformation may achieve this for parameters defined on the interval [0, 1].

Upon reflection, one might question whether the use of Frequentist style diagnostics on MCMC simulations of posterior distribtuions contradicts the underlying Bayesian modelling paradigm. The analysis of simulation output fits neatly within the Frequentist framework since each simulation can be seen as one of infinitely many similar simulations that could have been obtained under the same initial conditions. Properties of simulated distributions can be considered as fixed unknown values. In this situation it is the simulated data which is random and we are interested in confidence intervals: how likely is that future simulations will produce similar results? Contrast this with the application of the Bayesian philosophy to the modelling framework. Here we generally only have a single spatio-temporal dataset, the result of a single experiment, which is not easily repeated under exactly the same initial conditions. In this case, model parameters are best considered as random, and we are interested in which parameter values are best supported by the (fixed) data observed.

4.5 Statistical inference for MRFs

Reparameterization is an issue which had to be considered during implementation of the hierarchical models in the Bayesian approach to inference (Chapters 5–7). This issue is discussed in more detail in Section 4.5.1.

Several methods of estimating parameters in Markov random fields have been proposed in the literature, mostly variations on maximum likelihood. These are outlined in Sections 4.5.2 to 4.5.3.

Maximum Likelihood is not computationally feasible due to the intractable normalization constant. Asymptotic Maximum Likelihood (Pickard 1982) is based on an asymptotic Gaussian distribution for the energy function $\mathcal{H}$ and therefore the NC, as discussed in Section 4.5.2. The estimate of standard error of the parameter estimate needed improvement, so an Approximate Maximum Likelihood method was next considered in Section 4.5.2. This method is based on assessing clusters of cliques for saturation (all presences). All methods based on maximizing the log-likelihood suffer difficulties for a large lattice, since estimation of the standard errors require inversion of a matrix of second derivatives. The methods are

not appropriate if there is a high degree of interaction in the system. MLE approaches are discussed in Section 4.5.2.

Minimum χ^2 estimation (Section 4.5.3) uses a least squares approach to maximum likelihood, and is based on the log odds ratios of presence to absence given different neighbourhood configurations. This method also ignores spatial dependence between neighbourhoods when computing these average log odds ratios.

Coding, followed by Maximum Pseudo-Likelihood (Section 4.5.2), are two methods developed (Besag 1974) to overcome difficulties with the NC. Spatial dependence in the model is ignored by constructing a joint distribution of the lattice by multiplying together all the conditional distributions for each site.

4.5.1 Reparameterization

The parameterization of a model changes the shape of the likelihood surface, and so can affect the performance of any likelihood-based method, such as numerical integration/maximization of the likelihood and MCMC. The work of Kass & Slate (1992) develops diagnostics which indicate when joint and marginal posterior distributions deviate from Normality. They highlight that a major advantage of reparameterization is the resulting Normality of posteriors, leading to better estimation.

A few authors have addressed the reparameterization issue in the context of MCMC. Raftery & Lewis (1992) and Wakefield (1991) agree that the Gibbs sampler performs poorly for a parameterization which produces a highly correlated posterior distribution. This poor performance may arise from an inefficient sampler, which mixes slowly, taking a long simulation time to converge to the equilibrium distribution.

The literature provides four avenues to explore for reducing or eliminating parameter correlation: normalization, orthogonal transformation, centring via observed statistics, and centring via hierarchical modelling.

Hills & Smith (1992) explore transformations of the posterior distribution of parameters aiming towards Normality to eliminate this correlation.

Alternatively, it is possible to orthogonalise the parameter space (Muller 1994, for example) to remove correlations. However in higher dimensions this becomes increasingly difficult to implement. This is similar to the approach of Robert & Mengersen (1994) who investigate the use of perturbations when modelling location and scale parameters in mixture models.

Vines, Gilks & Wild (1994) attribute 'slow mixing' of the Gibbs sampler to lack of model identifiability. They solve this problem for a random effects model by transforming the random effects so that parameters are central on observed values of statistics.

Instead of centring on observed statistics, Gelfand, Sahu & Carlin (1994) and Gelfand, Sahu & Carlin (1995) propose centring parameters on other parameters using a hierarchical modelling approach. They show that the method of Vines et al. (1994) is deterministic rather than stochastic, and does not "borrow strength" within a layer of the model.

Instead of considering transformations of the model, Dellaportas (1995) considers a method of transforming the density during simulation MCMC, to achieve more efficiency and faster convergence. The method is related to importance sampling in that samples are obtained not from the density of interest but from a related function. Just as importance sampling is difficult to apply to MRF models, so is this method.

Within the Frequentist approach to modelling in this thesis (Chapter 3), the smoothed estimates of presence across sites did not allow full normalization of the model. Results

in Chapter 3 are not based on any transformation of the covariates q_k. Although centring with respect to observed statistics is not taken advantage of for the p_D parameters, they are modelled in a separate layer to the data and covariates, so this approach does achieve centring via modelling.

Initial work in Chapter 3 focuses on inference for the q parameters which represent the conditional probability of a success given presence and covariates. The difficulty with modelling parameters constrained to the unit interval is encountered mostly close to the boundaries of zero and one. In order to address this, the whole approach to modelling in Chapters 5 and 7 shifts emphasis from these q parameters, in fact expected means of the responses y, to the scale of a linear predictor η which is linked to the q parameters via a link function, such as the logit. Analysis in the Bayesian chapters proceeds based on the logit transformed α parameters, which are more Gaussian in distribution, and are not constrained to the unit interval.

The fully Bayesian approach requires the normalization constant of the autologistic model. After a theoretical treatment of estimation of the NC in Chapter 6, a four tier-model is investigated in Chapter 7. This approach goes some way towards normalizing the posterior densities of parameters, although the large size of the parameter space is prohibitive.

In preliminary results of Chapter 5 a link function may be defined to describe how the transformed response is related to the linear predictor and therefore the covariates using a Generalized linear model approach. In the *dingo* example, since the covariates simply comprise one factor with six levels, results can be equivalently derived on the untransformed scale of the response (as q_k), or alternatively on the transformed scale of the linear predictor (as α_k).

Centreing parameters via observed statistics is to some extent achieved for the spatio-temporal process by selecting the autologistic over the Ising parameterization. In the autologisitic model the canonical statistics relate naturally to statistics of interest, such as sum of presences, and sum of pairs of neighbouring presences.

Finally with this hierarchical model, we hope to achieve more identifiability of the spatio-temporal parameters, by placing them within a separate layer of the model. This contrasts with the 'flat' modelling approach of Preisler (1993), Heikkinen & Högmander (1994), and Denham & Mengersen (1999) where the spatio-temporal variation and effect of covariates share the same error structure.

Thus the parameterization issue is an important one, and has contributed to many of the modelling decisions in constructing the hierarchical model to be analyzed using a Bayesian approach. The frequentist approach addressed the parameterization issue to a very limited extent. Thus one of the advantages of the Bayesian approach is its flexibility.

4.5.2 Maximum Likelihood Estimation

A GRF belongs to the exponential family. If the parameterization equation (4.58) is used, then it follows directly (Cox & Hinkley 1974) that the energy function V is sufficient for canonical parameter θ. The log-likelihood is

$$\ell(\theta; z) = \theta^T V(z) - \log c(\theta) \tag{4.159}$$

The maximum likelihood estimator is obtained by differentiating and taking expectations (equivalent to the magnetization of equation (4.53) and equation (4.54)):

$$\frac{\partial \log c(\theta)}{\partial \theta} = \mathrm{E}[V(z)] \tag{4.160}$$

The critical points of the log-likelihood ℓ satisfy the normal equations

$$\frac{\partial \log c(\theta)}{\partial \theta} = V(z) \tag{4.161}$$

which has a unique solution $\theta = \widehat{\theta}$ which maximises ℓ.

The Hessian matrix is

$$-\mathrm{Cov}[V] = \frac{\partial^2 \log c(\theta)}{\partial \theta^2} \tag{4.162}$$

The Hessian corresponds to the specific heat quantity from the Ising model and thermodynamic contexts given on page 81. Note that using the Bragg-Williams or Bethe-Peierls approximations (Sections 4.3.5 and 4.3.5) to the log NC led to inaccuracies in estimating this quantity (Thompson 1988). The negative Hessian matrix is also the Jacobian of the transformation from θ to $\frac{\partial \log c(\theta)}{\partial \theta}$. Thus estimation of standard errors should be approached with caution.

Assuming that the edge sites are known, and then conditioning on these values, a maximum likelihood estimate for unknown parameters can theoretically be obtained. The unique MLE of unknown parameters θ is obtained by solving $K+1$ simultaneous normal equations corresponding to each of the $K+1$ equivalence classes of cliques for sampling window L

$$\mathrm{E}[V_k(z_L)] = V_k(z_L)\ k = 0, 1, \ldots K \tag{4.163}$$

where the conditioning on edge sites $\mathcal{N}(L)$ is subsumed into the notation for simplicity. The left hand side is the mean value of the kth sufficient statistic in the conditional distribution based on cliques of equation (4.18) for fixed parameters θ and edge sites $z_{\mathcal{N}(L_n)}$, and is also equal to (Possolo 1986*a*):

$$\frac{\partial \log c_L(\theta)}{\partial \theta_k} \tag{4.164}$$

Unfortunately the computation of the NC which requires a summation over the entire lattice, is not feasible. This problem is addressed later in Chapter 6 and applied to a Bayesian approach of inference in Chapter 7, although it could conceivably be applied to maximum likelihood as in Geyer & Thompson (1992). In addition computation of the variance estimates would require inversion of the Hessian, which for large lattices becomes computationally expensive.

Asymptotic Maximum Likelihood

As summarized by Possolo (1986*a*), the method developed by Pickard (1979) and Künsch (1983) used an asymptotic expression for the NC:

$$\frac{\log c_{L_n}(\theta)}{|L_n|} \to \Psi(\theta) \quad \text{as} \quad L_n \xrightarrow{\text{i.r.}} \mathcal{L} \tag{4.165}$$

where $\Psi(\theta)$ is a differentiable function of θ. The asymptotic MLE is then the solution of

$$\frac{\partial \Psi(\theta)}{\partial \theta_k} = \frac{V_k(z_{L_n})}{\gamma_k / u_k(L_n)} \tag{4.166}$$

and $u_k(L_n)$ is the number of summands in V_k (the number of cliques in the kth equivalence class), and

$$\gamma_k = \lim_{n\to\infty} \frac{u_k(L_n)}{|L_n|} \tag{4.167}$$

Specifically Ψ is derived by noting that the moment and cumulant generating functions, m_V and κ_V of the sufficient statistics $V_k(z)$ are related to the normalizing constant $c(\theta)$ as follows:

$$\begin{aligned} m_V(s;\theta) &= \mathrm{E}\left[\exp\left\{\sum_{k=0}^{K} s_k V_k(z)\right\}\right] \\ &= \frac{c(\theta+s)}{c(\theta)} \\ \kappa_V(s;\theta) &= \log m_V = \log c(\theta+s) - \log c(\theta) \end{aligned} \tag{4.168}$$

Pickard (1979) concludes that if the field is weakly interacting then

$$|L_n|^{-1/2}(V - \mathrm{E}[\,V\,]) \tag{4.169}$$

converges in distribution to a bivariate normal. Unless the field is weakly interacting, then both the MLE and asymptotic MLE are not consistent for θ (Pickard 1979).

Approximate Maximum Likelihood

A problem (Pickard 1982, Pickard 1979) with asymptotic MLE is that the location error for the standard error of $\tilde{\theta}$ (and thus also the acceptance and confidence regions) is at least as large as the standard error itself. So, unless $\frac{1}{N}\frac{\partial \log c(\theta)}{\partial \theta}$ can be estimated to within an accuracy of $O(N^{1/2})$, it is very difficult to attach a measure of reliability to the estimate $\tilde{\theta}$. According to Pickard (1976) likelihood inference for θ is "technically impossible" except for the classical Ising model with no single-site parameter and periodic boundary conditions where the partial derivatives of $\log c(\theta)$ converge at an exponential rate.

Even if maximum likelihood proceeds, conditional on unknown boundary site values (Strauss 1975) and effectively removing single-site parameter θ_0, the same problem is encountered. A large location error for the standard error of $\tilde{\theta}$ is again the problem, and cannot be avoided by fitting high-order cumulants.

The alternative described below was suggested by Pickard (1982) and is based on a reparameterization. Other authors have also investigated approximate maximum likelihood approaches Ogata & Tanemura (1981), Ogata & Tanemura (1984), Pickard (1982), Saunders, Kryscio & Funk (1979), Moran (1947), Bloemena (1964). Consider a clique measure of *saturation* with presence[14], which indicates presence at all sites within the clique:

$$s(C) = \mathrm{I}[\; z_i = 1 \;\forall i \in C\;] \tag{4.170}$$

or alternatively, for binary $z_i \in [0,1]$

$$s(C) = \prod_{i\in C} z_i. \tag{4.171}$$

Clique presence is not usually stationary within clique equivalence classes unless toroidal boundary conditions are used (Pickard 1982). With toroidal boundary conditions, the log

[14] called "clan interaction strengths" by (Pickard 1982)

normalizing constant $\log c(\theta)$ is distorted by the number of boundary sites $O(|N(L_n)|)$ (Hurst & Green 1960). However, we can consider the average clique saturation for cliques of order k over the entire lattice

$$\begin{aligned} s_k &= \frac{1}{|\mathcal{C}_k|} \sum_{C \in \mathcal{C}_k} \mathrm{E}[\, s(C)\,] \\ &= \frac{1}{|\mathcal{C}_k|} \mathrm{E}[\, V_k\,] \\ &= \frac{1}{|\mathcal{C}_k|} \frac{\partial \log c(\theta)}{\partial \theta_k} \end{aligned} \tag{4.172}$$

The second line obtains from noting that $V_k(z) = \sum_{C \in \mathcal{C}_k} \prod_{i \in C} z_i$. The third line simply applies equation (4.160). Due to a one-to-one relationship between θ and the first derivative of the log NC, one can use $s = \{s_k\}$ as a basis of estimation rather than θ. The MLE $\widehat{s}$ of s is

$$\widehat{s_k} = \frac{V_k(z)}{|\mathcal{C}_k|} \tag{4.173}$$

Furthermore Pickard (1982) showed that $\widehat{s}$ is unbiased and has minimum variance achieving the Cramer-Rao lower bound. Asymptotically, on a regular finite lattice, the standardised size of an equivalence class approaches the number of distinct orientations available to cliques in $\mathcal{C}_k$:

$$\frac{|\mathcal{C}_k|}{|L_n|} \to u_k \tag{4.174}$$

However, for asymptotics if clique saturations s are fixed as the sampling window increases, this requires adjustments to the original θ parameters. A complex system is then constructed (Pickard 1979) to achieve this based on well-known asymptotic behaviour of the log partition function in the Ising model. That is, in $\Re^2$, a result of Lebowitz (1974) gives the existence of differentiable function Ψ such that

$$\frac{1}{N} \log c(\theta) \to \Psi(\theta) \tag{4.175}$$

where Ψ is continuous and even in θ, and moreover matches the partial derivatives of the log partition function since

$$\frac{1}{N} \frac{\partial^{k+l} \log c(\theta)}{\partial \theta_k \partial \theta_l} \to \frac{\partial^{k+l} \Psi(\theta)}{\partial \theta_k \partial \theta_l} \tag{4.176}$$

occurs with certain regularity conditions for only some values of the parameter θ. (See Pickard (1979) for exact details.)

Maximum Pseudolikelihood

This method was developed by Besag (1974) as a method of overcoming the difficulties of estimating the normalizing constant in the likelihood for the auto-models. It is an extension of the earlier coding method which essentially viewed the lattice as two mutually independent portions under a first order neighbourhood system. A checkerboard of alternating black and red squares could be imposed on a lattice, indicating sites which were conditionally independent of each other. The coding estimator was then composed by averaging over all possible coding schemes. Maximum pseudo-likelihood generalized this idea

for general neighbourhood systems. This method is currently used for application papers using a Bayesian approach (Högmander & Møller 1995) due to its simplicity.

The true likelihood for the auto-logistic model is given in terms of the full joint distribution of the data and the parameters by:

$$\ell(\theta; z) = p(z|\theta) = \frac{\exp \theta^{\mathsf{T}} V(z)}{c(\theta)} \tag{4.177}$$

A likelihood based on conditional distributions could be constructed using

$$\ell(\theta; z) = \prod_{i=1}^{n} p(z_i|\{z_j;\ j < i\}). \tag{4.178}$$

Here the natural statistic $V(z)$ was defined in equation (4.59). A particular class of models which can easily be factored using the above relationship are called Markov Mesh Random fields (Abend, Harley & Kanal 1965, Qian & Titterington 1991). There is no simple way of writing this hierarchical conditional form of the likelihood in terms of the 'neighbourhood' conditional distributions from Section 4.2.6:

$$p(z_i|z_{\mathcal{N}(i)}, \theta) = \frac{\exp\{\theta^{\mathsf{T}} V_i(z)\}}{1 + \exp\{\theta^{\mathsf{T}} V_i(z)\}} \tag{4.179}$$

The pseudo-likelihood, however, has been devised since it can be expressed in terms of these neighbourhood conditional distributions

$$\wp\ell(\theta; z) = \prod_{i=1}^{n} p(z_i|\{z_j;\ j \in N(i)\}) \tag{4.180}$$

4.5.3 Minimum χ^2 Estimation

This method essentially examines the probability distribution of site values for every possible configuration of values at neighbouring sites. The larger the neighbourhood basis, the more numerous are the possible neighbourhood configurations.

In this section we present the method via two logical stages given in Derin & Elliott (1987), Chen (1988). Both stages rely on the stationarity and translation invariance of the local probability distribution which is therefore constant over all sites. Isotropic neighbourhood is assumed.

The first stage is the simple *least squares* solution to maximum likelihood (developed by Derin & Elliott (1987) for texture segmentation in image analysis) where the "error" in the energy function is estimated via a log odds ratio of absences to presences over constant neighbourhood configurations. One equation arises for each neighbourhood configuration. Computation can be simplified using only those neighbourhoods which occur most commonly on the observed lattice. With binary data, this is equivalent to a *Logit Estimation* method of Chen (1988).

The second stage compares the log odds of presence (or absence) comparing two different neighbourhood types *i.e.* cliques. Computation can be simplified by collapsing equivalence classes using symmetries.

From the literature, it appears that (independently) both Possolo (1986*a*) and Chen (1988) were inspired by Berkson (1949), Berkson (1956), Berkson (1980) to transfer the Minimum χ^2 method from Logistic Regression to this context. A debate revolving around

Berkson (1949) has raged for decades over the comparitive qualities of Maximum Likelihood *versus* Minimum χ^2 for Logistic Regression. Berkson (1980) claims that Minimum χ^2 is more accurate than Maximum Likelihood, which can in some contexts be equivalent to a form of χ^2 estimator.

The method hinges on the number of categories compared (as for χ^2 analysis), so it is essential to first reduce the number of clique equivalence classes via symmetry. With a first-order neighbourhood, there are $2^4 = 16$ distinct configurations, and with a second order one, there are $2^8 = 256$!

Consider marginal distributions of values within a neighbourhood. Label the possible configurations of the neighbourhood basis $\mathcal{N}$ as

$$\Omega(\mathcal{N}) = \{\eta_1, \eta_2, \ldots, \eta_m\} \tag{4.181}$$

where $m = 2^{|\mathcal{N}|}$ and imposing a sensible ordering $\{1, 2, \ldots m\}$. Estimate the observed marginal log odds of each map value z_i occuring with a given neighbourhood $z_{\mathcal{N}(i)} = \eta_m$ by

$$\widehat{P}\{z_i = z_i, z_{\mathcal{N}(i)} = \eta_m\} = \sum_{i \in L_n} I[z_i = z_i \quad \text{and} \quad z_{\mathcal{N}(i)} = \eta_m]. \tag{4.182}$$

for each $\eta_m \in \Omega(\mathcal{N})$. Derin & Cole (1986) simplify this to

$$\widetilde{P}\{z_i = z_i, z_{\mathcal{N}(i)} = \eta_m\} = \sum_{i \in L_n} I[z_i = z_i,\ P(z_i = 0 \,|\, z_{\mathcal{N}(i)}) = P(z_i = 0 \,|\, \eta_m)] \tag{4.183}$$

for a reduced number of simplified neighbourhood configurations $\{\eta_1, \eta_2, \ldots \eta_{\widetilde{m}}\}$.

Using the property that z is a MRF with the attending positivity and translation invariance conditions, Possolo (1986*a*) derive the following relationship:

$$\log \frac{\Pr\{z_i = 1 \,|\, z_{\mathcal{N}(I)} = \eta_m; \theta\}}{\Pr\{z_i = 0 \,|\, z_{\mathcal{N}(I)} = \eta_m; \theta\}} = \theta^T W = \sum_{k=0}^{K} \theta_k \omega_{km} \qquad \forall i \in I \tag{4.184}$$

where $\omega_{0m} = 1$ and $\omega_{km} = \sum_{C \in \mathcal{C}_k} I[(\eta_m)_j = 1 \forall j \in C]$ counts the number of k-order cliques C on the lattice which match the corresponding portion of the mth neighbourhood configuration η_m, containing the centre site of η_m, and contain all presences.

A finite sample adjustment needs to be used for the sample block probabilities to ensure the denominator in equation (4.184) is nonzero

$$P_{\text{adj}}(\cdot) = P(\cdot) + \epsilon \tag{4.185}$$

where ϵ is a very small number. In the unsimplified case ($\widehat{P}$), let $\epsilon = 2^{-(M+1)}$ and in the simplified case ($\widetilde{P}$), let $\epsilon = 2^{-(\widetilde{M}+1)}$.

Now define the sample average of the number of sites in the sampling window L_n with neighbourhood configuration η_m.

$$\widehat{P}_n(z, \eta_m) = \frac{1}{|L_n|} \sum_{i \in L_n} \mathrm{I}\left[\, z_i = z, z_{\mathcal{N}(i)} = \eta_m \,\right] \tag{4.186}$$

Theorem 4.4 *Possolo (1986*a*) The sample log ratio*

$$P_{mn} = \log \frac{\widehat{P}_n(1, \eta_m)}{\widehat{P}_n(0, \eta_m)} \tag{4.187}$$

is strongly consistent for the linear combination $\theta^T W$ as the sampling window $L_n \xrightarrow{i.r.} \mathcal{L}$.

The proof is given in Possolo (1986*a*) and appeals to ergodicity deriving from the coherence of the conditional distributions.

A Least Squares Solution

Possolo (1986*a*) proposes to estimate $\hat{\theta}_n$ using Least Squares via

$$\Sigma_n^{-1/2} V_n = \Sigma_n^{-1/2} W\theta + \xi_n \tag{4.188}$$

where Σ_n is the $m \times m$ variance matrix of R_{mn}, and $\xi_{nm} \xrightarrow{\text{a.s.}} 0$ as $L_n \xrightarrow{\text{i.r.}} \mathcal{L}$. Consistency is obtained due to the previous theorem.

Replacing Σ_n by a sample estimate is necessary in practice, and only affects the precision and not the validity of the LSE $\hat{\theta}$ (Possolo 1986*a*).

Variance estimtaes can be derived by assuming that the sample probabilities $\widehat{P}_n(1, \eta_m)$ and $\widehat{P}_n(0, \eta_m)$ can be viewed as joint trinomial frequencies. Estimate Σ_n by $diag(\widehat{\sigma}_{nm})$ with $\widehat{\sigma}_{nm}$ being a Taylor series approximation to the variance of R_{mn}:

$$\begin{aligned}
\text{Var}[\, R_{mn}\,] \;\simeq\; & \frac{1}{|L_n|}\left\{\frac{1-p_{m1}}{p_{m1}} + \frac{1-p_{m0}}{p_{m0}} + 2\right\} \\
+ \; & \frac{1}{|L_n|^2}\left\{\frac{2p_{m1}-1}{p_{m1}} + \frac{2p_{m0}-1}{p_{m0}} + \frac{(1-p_{m1})(1-p_{m0})}{2p_{m1}p_{m0}}\right. \\
+ \; & \left.\frac{(1-p_{m1})(3p_{m1}-1)}{2p_{m1}^2} + \frac{(1-p_{m0})(3p_{m0}-1)}{2p_{m0}^2}\right\} \\
+ \; & O(|I_n|^{-3})
\end{aligned} \tag{4.189}$$

With a large enough sampling window, the sample log-ratios R_{mn} behave well, according to Possolo (1986*a*). Quoting a result from Dobrushin (1968), Theorem 5, for discrete MRFs with finite-range, the process is exponentially mixing. This can be used to advantage in improving the variance estimates.

Neighbourhood basis estimation

Possolo (1986*a*) suggests following a procedure similar to subset regression to choose the optimal neighbourhood basis according to some criterion of goodness-of-fit and parsimony, *e.g.* AIC. It is important to note that parsimony is an especially desirable feature for two reasons. The number of parameters explodes exponentially as the neighbourhood basis expands in size, posing computing restraints. Some of the counts $\widehat{p}_n(z, \eta_m)$ may be zero if large neighbourhoods are used, resulting in indeterminate sample log-ratios. Possolo (1986*a*) suggests a finite sample adjustment, such as adding $\frac{1}{2}|L_n|^{-1}$ to each count before computing the log-ratios *c.f.* Haldane-Anscombe estimator. Some of these drawbacks can be avoided by taking advantage of any symmetry in the neighbourhood structure, *e.g.* if two rows of W are identical, *e.g.* $\omega_{km_1} = \omega_{km_2}\ \forall k$, then so are $\widehat{p}_n(z, \eta_{m_1})$ and $\widehat{p}_n(z, \eta_{m_2})$.

4.6 Discussion

The initial exploratory data analysis (Section 3.2) and frequentist approach to analysis (Section 3.3) showed that the dingo data exhibited evidence of both temporal and spatial dependence. Following the hierarchical modelling structure of depending on covariates and an underlying spatio-temporal process. Of the binary spatial models considered in the review in Section 2.4, Markov random field (MRF) models showed the most potential as the underlying spatio-temporal model. An advantage of MRFs is their dual representation as both a joint probability density describing the macroscopic behaviour of all sites

contemporaneously, and equivalently as the local probability density describing the microscopic relationship between sites within a neighbourhood. This hierarchical model differs from, but can also borrow from, those popular in many image analysis applications (Besag 1986, Dubes & Jain 1989, Weir & Pettitt 1999) to describe black and white images blurred by noise.

Once the overall hierarchical modelling structure had been selected, a suitable MRF was selected, as well as an accompanying approach to inference.

Essentially the parameterization choice for a binary Markov random field—Ising *versus* Autologistic—depends on the ease with which the model can be interpreted (Gelman et al. 1995), or computational efficiency issues (Hills & Smith 1992, Kass & Slate 1992, Vines et al. 1994). Interpretive ease depends to a large extent on the application. The autologistic model emphasises the contrast between presence and absence whereas the Ising model emphasises similarity and dissimilarity. The autologistic dependence parameter θ influences the likelihood of a pair of 'present' sites being adjacent whereas the Ising analogue β regulates the balance between the number of similar and dissimilar pairs.

The Ising parameterization does achieve centring for parameters: prevalence represents the deviation from an overall balance of 50-50 presence-absence; and dependence reflects the balance between the similar and dissimilar pairs. Computational benefits arise from dealing with sums centred around zero; although some values of the Ising parameters nevertheless lead to sums of the same order of magnitude as the autologistic version. The autologistic parameterization lends itself more readily to an extension with covariates. Its relationship with the Spatial Linear Models or Spatial ARIMA models (Cressie 1993) is more easily recognized, and has been utilized in non-hierarchical spatial models (Preisler 1993, Wu & Huffer 1997).

It is for this last reason as well as the presence/absence interpretation that I focus on the autologistic parameterization. I will use the shorthand AL(θ) to denote the autologistic model with parameter θ.

Other modelling choices include whether isotropy is to be assumed or not, and the order of neighbourhood to be considered. Generally, it is more flexible to allow non-isotropy in the neighbours, although this requires extra parameters and with sparse data can lead to identifiability problems. Finally, in this thesis I will consider only first order neighbourhoods due to computing constraints although the results in this section extend to larger neighbourhoods.

Theoretical results available for the Ising model in statistical physics highlight critical portions of the parameter space for the spatio-temporal interaction parameters θ_1, θ_2. Values over approximately 1.38 could lead to difficulties with simulation and therefore inference. The existence of this critical value with the Ising/autologistic model is what distinguishes it from its continuous counterpart, the auto-Gaussian model. For parameter values beyond a critical point, presence and absence on the lattice begin to behave co-operatively, a form of long-range dependence. With a Gaussian Markov random field, these critical values do not occur. Hence the Ising/autologistic model is suitable for applications where it is thought that this co-operative behaviour might exist between lattice sites.

A literature review of existing methods of inference for MRFs revealed several methods which have various advantages and disadvantages. Variations on maximum likelihood suffer from difficulty with estimating the standard errors of estimates. Pseudo-likelihood ignores dependence between conditional distributions used to construct the pseudo log-likelihood. Minimum χ^2 needs careful selection and assessment of categories of cliques used to estimate log odds ratios of presence to absence.

It was therefore thought worthwhile to consider a Bayesian approach in order to overcome these obstacles. The posterior distributions of parameters are difficult to obtain analytically for this model, so simulation methods were considered. No method of obtaining independent samples from MRFs is available in the literature, however, it is relatively easy to obtain dependent simulations from MRF models using the method of Markov Chain Monte Carlo (MCMC).

In this thesis I settled on the simple Metropolis-Hastings MCMC sampler for the autologistic model with parameter θ incorporating prevalence component θ_0. This was a simple choice, and avoided augmenting the system to consider simulation of V_0 or T in addition to the site values z. More complex simulation methods will be considered in future work.

For computations, I used a single long run to avoid the bias and convergence problems of shorter runs. Preliminary work compared results from multi-starting at different starting values. Integrated autocorrelation time was estimated using the truncated periodogram estimator given in Equations 4.148 and 4.149, with a value of $c = 3$. Run length was determined based on IACT using Equation 4.156.

The hierarchical Bernoulli - autologistic model framework is developed and applied to the *Dingo* case study in the next chapter, Chapter 5.

Chapter 5

Bayesian hierarchical models for 2D binary lattice data with underlying spatial dependence

Contents

5.1 Introduction 124
5.2 Bayesian hierarchical model 124
5.2.1 Three tier hierarchical model 125
5.2.2 Four tier hierarchical model 128
5.3 Inference 128
5.3.1 Initialization of θ 129
5.4 Computations: MCMC Design 130
5.4.1 Sampler for unknown presence/absence z_{sr} 130
5.4.2 Sampler for success/failure probabilities q_k 131
5.5 Pilot simulation experiment 132
5.5.1 Initialization 133
5.5.2 Results from pilot experiment 134
5.5.3 Posterior density of q_k 134
5.5.4 Posterior density of z_{sr} 141
5.5.5 Checking convergence 146
5.6 Proposal experiment 148
5.6.1 Conclusion 151
5.7 Bayesian model choice 154
5.7.1 Bayes factors 154
5.8 Discussion 156

5.1 Introduction

> The chapter of knowledge is a very short, but the chapter of accidents is a very long one.
> *- Lord Chesterfield: letter to Solomon Dayrolles, 16 February 1753*

Building on the foundation of Chapter 4 this chapter begins an investigation, spanning the next three chapters, into Bayesian modelling and inference of spatial presence/absence data with ambiguous zeroes. This contrasts with the frequentist approach employed in Chapter 3; its relative benefits and drawbacks were discussed in detail in Section 3.4. The Bayesian approach taken in this chapter addresses the major difficulties, namely, in the way that spatial dependence was modelled, interpreting results, and estimating standard errors.

A hierarchical modelling approach can be implemented within a Bayesian paradigm, with many benefits (Gelman et al. 1995) in this context. Spatio-temporal dependence can be modelled via the three-parameter autologistic model, a binary Markov random field (MRF) which was identified as potentially suitable in Section 2.4. Full posterior distributions for parameters can be derived to give a more complete picture of accuracy than point and standard error estimates. Computation supporting inference for MRFs can be achieved using Markov Chain Monte Carlo (MCMC) as demonstrated later in Section 5.4. These benefits are attained at the cost of additional computational effort.

Both a three-tier model and a four-tier hierarchical model are investigated in this thesis. These are defined in Section 5.2. I begin by developing inference for the three-tier model in this chapter. I address the difficult Normalization constant problem in Chapter 6, a necessary prerequisite for proceeding with the four-tier model in Chapter 7.

Chapter 4 introduced the theoretical framework and statistical tools (Section 4.4) underpinning this approach. The Bayesian approach and its implementation using Markov Chain Monte Carlo (MCMC) techniques are outlined for this particular problem. With reference to the literature review in Section 4.4, various methods of constructing these Markov Chains are presented, including hybrid methods. Practical issues of implementing MCMC in this context are addressed.

Section 5.5 applies this theory to the *dingo* case study investigating the attractiveness of various chemicals to dingos. Particular attention is paid to diagnosing whether the MCMC chains have converged to equilibrium in Section 5.5.5. Implementation of the MCMC computational method requires selection of proposal and prior distributions for parameters. Sensitivity of analysis to the choice of proposal distributions (Section 5.6) and of prior distributions (Section 5.3.1) is therefore investigated.

Finally, an *ad hoc* approach to choice of parameters in the prior for spatio-temporal dependence in the three tier model is investigated in Section 5.7.

5.2 Bayesian hierarchical model

Consider a multivariate binary random variable observed on a spatio-temporal grid $y = \{y_{vsr} : v = 1, \ldots, V; s = 1, \ldots, S; r = 1, \ldots, R\}$, with v indexing the variable, s the spatial location, and r the time stamp. Recall that in the *dingo* case study presented in Chapter 3, s and t denote spatial and temporal position represented by site along the transect and day of sampling. There were bivariate responses measured at every site, corresponding to each location within a site, so index v denotes this location within a site. An extension

not considered in this thesis would allow y_{vsr} and $y_{v'sr}$ to be correlated instead of assuming dependence, as is done in the thesis.

The sample space is $y_{vsr} \in \{0,1\}$ and overall $y \in \Omega = \{0,1\}^{VSR}$. Here one and zero can denote, for instance, presence and absence, success and failure or black and white. Without loss of generality the success and failure description for y shall be used.

Section 5.2.1 defines the three-tier model, and section 5.2.2 extends this definition to the four-tier model.

5.2.1 Three tier hierarchical model

In Figure 5.1, a graphical representation of the full three-tier model is given. This highlights the conditional independence relationships between data y components given covariates x, coefficients α and spatio-temporal process z. This figure also displays the conditional dependence structure within z. These relationships are described in more detail in the following paragraphs, and explicitly defined in equation (5.1)-equation (5.7).

We introduce an underlying spatio-temporal presence/absence process $z = \{z_{sr} : s = 1,\ldots,S;\ r = 1,\ldots,R\}$ with $z_{sr} \in \{0,1\}$ where zero denotes absence and one denotes presence. Presence for a component of z signals that it is possible to observe a success for the corresponding component of y. So there exists v such that $y_{vsr} = 1$ implies $z_{sr} = 1$. Absence for a component of z necessarily requires failure for the corresponding component of y. That is, $z_{sr} = 0$ implies $y_{vsr} = 0$ for all v.

We model the conditional distribution of y given z like a Generalized linear model (GLM) of McCullagh & Nelder (1993), with $\mathrm{E}[y] = \mu$ and link function $g(\mu) = \eta$ linking the mean to a linear predictor $\eta = X^{\mathsf{T}}\alpha$. Here X is a matrix of covariates or a design matrix and $\alpha = [\alpha_1,\ldots,\alpha_k]^{\mathsf{T}}$ is an unknown vector of coefficients. In the *dingo* case study analysis that follows I choose a logit link function

$$\eta = g(\mu) = \text{logit}\,(\mu) = \log\left(\frac{\mu}{1-\mu}\right) \quad \text{and} \quad \mu = g^{-1}(\eta) = \frac{e^{\eta}}{1+e^{\eta}}$$

although a number of other link functions suitable for binary data may also be appropriate.

Note that

$$g(\mu_{vsr}) = \eta_{vsr} = \sum_{k=1}^{K} X_{k,vsr}\alpha_k$$

where $X_{k,vsr}$ denotes the element of the $K \times VSR$ covariate/design matrix X in the kth row and in the column indexed by tuple v,s,r. In the case where X is a design matrix comprising all 0s and 1s, with a single 1 in each column, then this simplifies to $g(\mu_{vsr}) = \eta_{vsr} = \alpha_k$ where k is such that $X_{k,vsr} = 1$ for a single k and $X_{k',vsr} = 0$ for all other $k' \neq k$. A further simplification is that $\mu_{vsr} = g^{-1}(\eta_{vsr}) = g^{-1}(\alpha_k)$. For compatibility with earlier work (Pettitt & Low Choy 1999), the notation $\tau_{vsr} = k$ indicates that $X_{k,vsr} = 1$ and $X_{k',vsr} = 0$ for all $k' \neq k$; also define $q_k = \mu_{vsr}$ when $\tau_{vsr} = k$. In any case, inference could equally proceed based on q, μ, η or α since there is a direct one-to-one and therefore invertible logical relationship between each pair. The advantages of proceeding based on α are those derived from the logit transformation (Cox 1970), as per Section 4.5.1. The advantages of proceeding based on q are interpretative, and consistency with Chapter 3 and Pettitt & Low Choy (1999).

Assume that the components of y are conditionally independent given z, that is, all spatio-temporal dependence is introduced via the underlying spatio-temporal process or

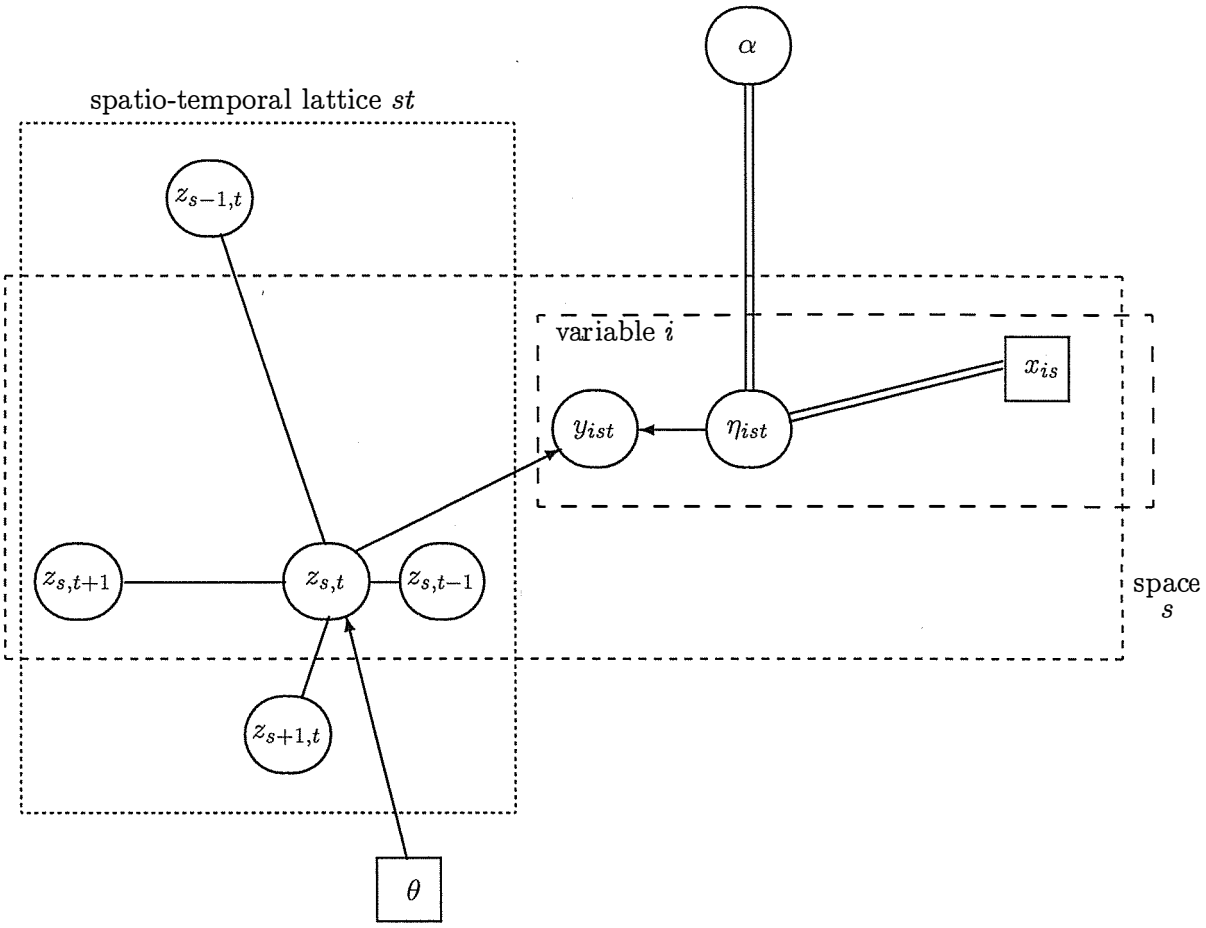


Figure 5.1: Relationship between variables in initial hierarchical model for *Dingo* case study, with fixed θ. The key to reading a Dynamic Acyclic Graph such as this is as follows. Squares enclose fixed parameters and circles enclose stochastic parameters in the model. Arrows indicate stochastic dependence of target variable on source variable; double lines indicate deterministic relationships. The plates (large dashed rectangles) enclose groups of dependent variables.

the covariates. Then the likelihood factorises as

$$p(y \mid q, z) = \prod_{v=1}^{V} \prod_{s=1}^{S} \prod_{r=1}^{R} p(y_{vsr} \mid q_{\tau_{vsr}}, z_{sr})$$

The joint distribution of y, z and q can be written in terms of the likelihood and distributions of q and z:

$$\begin{aligned} p(y, z, q \mid \theta) &= p(y \mid z, q)\, p(z, q \mid \theta) \\ &= p(y \mid z, q)\, p(z \mid \theta)\, p(q) \\ &= p(y \mid z, q)\, p(z \mid \theta) \prod_{k=1}^{K} p(q_k). \end{aligned} \tag{5.1}$$

I shall refer to this as the *three-tier hierarchical model.* By construction, parameters z and q are independent. We note that q is determined by α in a one-one monotonic increasing and deterministic (logical) relationship, and so the prior for q is determined directly from those for α or η. In turn the prior for q can be factorized into a product of priors for each component under the assumption of independence. A non-informative prior for the coefficients q_k is suitable unless more specific information is available:

$$q_k \sim \text{IID Uniform}(0, 1) \quad k = 1, \ldots, K.$$

Since y comprises conditionally independent binary variates, a Bernoulli distribution is appropriate for each component y_{vsr} given z_{sr}, with probability of success being given by $q_{\tau_{vsr}}$:

$$p(y_{vsr} | q_{\tau_{vsr}}, z_{sr}) = \begin{cases} (q_{\tau_{vsr}})^{y_{vsr}} (1 - q_{\tau_{vsr}})^{(1-y_{vsr})}, & z_{sr} = 1 \\ 1 - y_{vsr}, & z_{sr} = 0 \end{cases} \tag{5.2}$$

The prior $p(z \mid \theta)$ for the spatio-temporal process z given parameter θ can be modelled by a binary Markov random field, as discussed in Chapter 3, *i.e.*

$$p(z \mid \theta) = \frac{\exp\{\theta^R V(z)\}}{c(\theta)}. \tag{5.3}$$

An example of a binary MRF is the three parameter autologistic (Section 4.3) where

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix} \quad \text{and} \quad V(z) = \begin{bmatrix} \sum_{sr} z_{s,r} \\ \sum_{sr} z_{sr} z_{s+1,r} \\ \sum_{sr} z_{sr} z_{s,r+1}. \end{bmatrix} \tag{5.4}$$

This equation is equivalent to equation (4.62) although the spatio-temporal index i is separated into two components s and r to aid interpretation of the *dingo* case study. Here θ_0 relates to prevalence (fugacity of page 91), θ_1 describes dependence in space and θ_2 dependence in time. An equivalent formulation for the autologistic is provided by the conditional distribution of z_{sr} given all other spatio-temporal sites $\{-sr\} = \{(s', r') : \text{not } (s' = s \text{ and } r' = r)\}$ with

$$\begin{aligned} h_{sr}(z_{sr}, z_{-sr}, \theta) &= \exp\{z_{sr}[\theta_0 + \theta_1(z_{s-1,r} + z_{s+1,r}) + \theta_2(z_{s,r-1} + z_{s,r+1})]\} \\ p(z_{sr} \mid z_{-sr}, \theta) &= \frac{h_{sr}(z_{sr}, z_{-sr}, \theta)}{h_{sr}(0, z_{-sr}, \theta) + h_{sr}(1, z_{-sr}, \theta)}. \end{aligned} \tag{5.5}$$

Hence

$$\begin{aligned} h_{sr}(z_{sr}, z_{-sr}, \theta) &= \log \frac{p(z_{sr} = 1 \mid z_{-sr}, \theta)}{1 - p(z_{sr} = 1 \mid z_{-sr}, \theta)} \\ &= \text{logit}\,(p(z_{sr} = 1 \mid z_{-sr}, \theta)) \end{aligned} \tag{5.6}$$

As the prior probability of a presence increases, so must the log odds ratio of presence increase. The expression h_{sr} can be viewed as the prior log odds of presence given information on neighbours. The larger the value of θ_1, the larger the contribution to h_{sr} if spatial neighbours are present. So θ_1 represents the effect of spatial neighbours on presence. Similarly θ_2 represents the effect of temporal neighbours on presence. The parameter θ_0 represents the odds of a visit if there were no visits to neighbouring sites at the same time period, or the same site on neighbouring time periods. Also, the larger the contribution from θ_0 in comparison to the spatio-temporal θ_1, θ_2, the less the emphasis on the observed data collected about visits to neighbouring spatio-temporal site combinations.

5.2.2 Four tier hierarchical model

Suppose that θ is random, and can be assigned a probability density which depended on parameter vector ϕ. Then the joint distribution in equation (5.1) can be extended to include another layer

$$\begin{aligned} p(y, z, q, \theta \mid \phi, X) &= p(y \mid z, q, X)\, p(z \mid \theta)\, p(\theta \mid \phi) \prod_k p(q_k) \\ p(z \mid \theta) &= \frac{h(z, \theta)}{c(\theta)} \end{aligned} \tag{5.7}$$

Here ϕ is assumed independent of other parameters q and the covariates X, and z depends on ϕ only through θ. This extended model shall be referred to as the *four-tier hierarchical* model.

5.3 Inference

Within the Bayesian formulation, inference focuses on deriving posterior marginal distributions for parameters. From the posterior distribution many descriptive statistics may be derived. Central tendency can be described by the mean or the median and variability by quantiles or credibility intervals, for example. Posterior distributions are directly related to the joint distribution by Bayes' theorem Let u represents observed data and λ represents a vector of unknown random parameters. The posterior distribution for a component λ_k extracts from this expression only those terms which involve that component, all other terms are absorbed into the constant of proportionality. So

$$p(\lambda_k \mid u) \propto p_k(u \mid \lambda)\, p_k(\lambda).$$

So $p_k(\cdot)$ is generic notation for the function derived from $p(\cdot)$ which only involves terms containing λ_k.

First consider the posterior conditional distribution of the spatio-temporal absence/presence process. It is necessary to be aware of those elements of the latent spatio-temporal process where z is unknown and those elements where z is known. Cases where z is known arise

because at least one success was observed at that spatio-temporal site, that is, $y_{vsr} = 1$ for some v. The posterior conditional distribution of z_{sr} can be derived

$$p(z_{sr} \mid \ldots) \propto p(z_{sr} \mid z_{-sr}, \theta) \prod_{v=1}^{V} p(y_{vsr} \mid z_{sr}, q_{\tau_{vsr}}) \tag{5.8}$$

Similarly the posterior conditional distribution of q_k (or equivalently α_k) is given by

$$p(q_k \mid \ldots) \propto p(q_k) \prod_{vsr\,:\,\tau_{vsr} \neq 0} p(y_{vsr} \mid z_{sr}, q_{\tau_{vsr}}) \tag{5.9}$$

Inference for this model using analytic or numerical approaches, such as empirical Bayes (Gelman & Meng 1996, for example), is restricted by the dimensionality and simultaneous nature of the dependence in the underlying spatio-temporal process $p(z \mid \theta)$. When the three tier model of equation (5.1) is used for $p(y, z, q \mid \theta)$, then an approach can be used which avoids use of the joint distribution $p(z \mid \theta)$ by evaluating ratios of the full conditional distribution $p(z_{sr} \mid z_{-sr}, \theta)$. When the four tier model of equation (5.7) is used for $y\, z\, q\, \theta \mid \phi$, then evaluation of the full joint distribution, including its normalization constant $c(\theta)$ cannot be avoided.

In this chapter I begin by developing inference for the three-tier model of equation (5.1). The fourth tier can be considered to a limited extent in the three-tier context by using Bayes factors (Section 5.7). Full implementation of the fourth-tier requires evaluation of ratios of Normalization constants for the prior for spatio-temporal dependence. I address the difficult Normalization constant problem in Chapter 6, and then proceed with inference for the four-tier model of equation (5.7) in Chapter 7.

Inference is supported by application of the Markov chain Monte Carlo method (Metropolis et al. 1953, Gilks et al. 1996) described in Section 4.4 which is suitable for simulating dependent samples from these posterior distributions.

5.3.1 Initialization of θ

The three-tier model of equation (5.1) conditions on the parameter θ. Sensitivity analysis via Bayes Factors in Section 5.7 can be used to investigate comparison of results for different θ values. Later these parameters are explicitly modelled using hyperpriors in the four-tier model of equation (5.7). Particular values of $\theta = (\theta_0, \theta_1, \theta_2)$ reflect various types of spatio-temporal behaviour.

Denote by $\hat{p}_{h,v}$ the probability of a presence given that neighbours were present at h horizontally-adjacent sites, and neighbours were present at the same site on v vertically-adjacent sites. Thus h is the number of horizontal neighbours, and v the number of vertical neighbours. On the *dingo* two-dimensional spatio-temporal lattice h will represent the number of spatial neighbours on the lattice, and v the number of temporal neighbours. In the previous notation of Chapter 3 this is

$$\hat{p}_{h,v} = \Pr\left(y_{sr} = 1 | D_{s-1,r} + D_{s+1,r} = h; D_{s,r-1} + D_{s,r+1} = v\right) \tag{5.10}$$

corresponding to general notation of Chapter 4 of

$$\hat{p}_{h,v} = \Pr\left(y_{vsr} = 1 | z_{s-1,r} + z_{s+1,r} = h; z_{s,r-1} + zD_{s,r+1} = v\right).$$

One can assign the probability of a presence, $\hat{p}_{0,0}$ when there are no horizontal or vertical neighbours, i.e. $s = r = 0$. Then θ_0 can be obtained from the prior for dingo presence D:

$$\theta_0 = \text{logit}\,(\hat{p}_{0,0}) \tag{5.11}$$

Suppose $\theta_1, \theta_2 \geq 0$, which means that more presences in the neighbourhood tend to increase (rather than decrease[1]) the chances of a presence at the central position, then h_{sr} is maximised when all neighbours are present, *i.e.* $s = r = 2$.

$$\theta_1 + \theta_2 = \frac{1}{2}\left(\text{logit}\,(\hat{p}_{2,2}) - \text{logit}\,(\hat{p}_{0,0})\right) \tag{5.12}$$

One way of choosing θ_1 and θ_2 is by assigning $\hat{p}_{2,0}$, the probability of presence when all horizontal but no vertical neighbours are present:

$$\begin{aligned} \theta_1 &= \frac{1}{2}\left(\text{logit}\,(\hat{p}_{2,0}) - \text{logit}\,(\hat{p}_{0,0})\right) \\ \theta_2 &= \frac{1}{2}\left(\text{logit}\,(\hat{p}_{2,2}) - \text{logit}\,(\hat{p}_{2,0})\right) \end{aligned} \tag{5.13}$$

Another intuitive choice is to set a value for the ratio of horizontal to vertical effect $\rho = \theta_1/\theta_2$. Then θ_1, θ_2 and ρ may be expressed in terms of $\theta_1 + \theta_2$ obtained in equation (5.12).

$$\begin{aligned} \theta_1 &= \frac{\rho}{\rho + 1}(\theta_1 + \theta_2) \\ \theta_2 &= (\theta_1 + \theta_2) - \theta_1 \end{aligned} \tag{5.14}$$

5.4 Computations: MCMC Design

Simulation from the posterior distributions $p(z_{sr} \mid \ldots)$ and $p(q_k \mid \ldots)$ can be achieved simultaneously by a hybrid Markov chain Monte Carlo algorithm as per Section 4.4.4. The algorithm is hybrid since different samplers are tailored to simulation for each component of the posterior: section 5.4.1 details simulation from z_{sr}; and section 5.4.2 describes simulation for α_k.

5.4.1 Sampler for unknown presence/absence z_{sr}

Due to the equivalence between the full conditional $p(z_{sr} \mid z_{-sr}, \theta)$ and joint $p(z \mid \theta)$ formulations (Section 4.2.6) the posterior distribution of z_{sr} only involves the local formulation. One of the most direct MCMC samplers is a Gibbs sampler, which samples directly from the full posterior distribution of the parameter (Section 4.4.3). I derive the full posterior distribution of z_{sr} to show that the Gibbs sampler may be applied here. Using equation (5.2) and equation (5.3) we obtain

$$\begin{aligned} p(z_{sr} \mid \ldots) &\propto p(z_{sr} | z_{-sr}, \theta) \prod_{v=1}^{V} p(y_{vsr} \mid z_{sr}, q_{\tau_{vsr}}) \\ &= \frac{h_{sr}(z_{sr}, z_{-sr}, \theta)}{h_{sr}(0, z_{-sr}, \theta) + h_{sr}(1, z_{-sr}, \theta)} \end{aligned}$$

[1] In the *Dingo* case study, if dingos are more systematic and tend not to cover area they covered the previous day, for instance, then θ_2 could perhaps be negative. However, we do not consider this option as it was considered unlikely by biological experts.

$$\times \prod_v (z_{sr}q_{\tau_{vsr}} + (1 - z_{sr}))^{y_{vsr}}(1 - z_{sr}q_{\tau_{vsr}})^{(1-y_{vsr})} \tag{5.15}$$

Now z_{sr} unknown implies $y_{vsr} = 0$, for all v. This simplifies the first term in the product to be $\prod_v (1 - z_{sr}q_{\tau_{vsr}})^{(1-y_{vsr})}$. The second term in the product evaluates to one since $y_{vsr} = 0$, for all v whenever $z_{sr} = 0$. Now since z_{sr} is binary we can evaluate the odds ratio to eliminate the denominator in equation (5.15)

$$\frac{p(z_{sr} = 1 \mid \ldots)}{p(z_{sr} = 0 \mid \ldots)} = \frac{h_{sr}(1, z_{-sr}, \theta)}{h_{sr}(0, z_{-sr}, \theta)} \prod_v (1 - q_{\tau_{vsr}}) = \kappa_{sr}, \text{ say.} \tag{5.16}$$

But

$$\kappa_{sr} = \frac{p(z_{sr} = 1 \mid \ldots)}{1 - p(z_{sr} = 1 \mid \ldots)} \quad \text{so} \quad p(z_{sr} = 1 \mid \ldots) = \frac{\kappa_{sr}}{1 + \kappa_{sr}} = (\kappa_{sr}^{-1} + 1)^{-1}. \tag{5.17}$$

Thus the full posterior distribution of z_{sr} is available. It is easily sampled from since

$$z_{sr} = \begin{cases} 1 & \text{with probability } (\kappa_{sr}^{-1} + 1)^{-1} \\ 0 & \text{with probability } (\kappa_{sr} + 1)^{-1} \end{cases}$$

Hence the posterior distribution for z_{sr} can be easily evaluated for all z_{sr}. Thus the Gibbs sampling algorithm is most appropriate for simulating these components. The Gibbs sampling algorithm for z_{sr} is

Step 1 Initialize z_{sr} to $z_{sr}^{(0)}$.

Step 2 Repeat the following steps for iterations $n = 1, \ldots, N$

Step 2a Generate uniform variate $U \sim \text{Uniform}(0, 1)$.

Step 2b Evaluate κ_{sr} as per equation (5.16) using the most up-to-date values $z_{-st}^{(n-1)}$ and $q_{\tau_{vsr}}^{(n-1)}$ for v, s, r not updated yet and $z_{-st}^{(n)}$ and $q_{\tau_{vsr}}^{(n)}$ for v, s, r already updated.

Step 2c If $U \leq (\kappa_{sr}^{-1} + 1)^{-1}$ then set $z_{sr}^{(n)} = 1$ else set $z_{sr}^{(n)} = 0$.

Step 3 Check convergence and efficiency of algorithm (referring to Section 4.4.7). If necessary, repeat Step 2 until convergence to equilibrium is achieved with sufficient samples.

5.4.2 Sampler for success/failure probabilities q_k

The posterior distribution for q_k is not easily specified. However, ratios of posterior distributions are easily evaluated. Therefore the Metropolis-Hastings sampler is appropriate (Section 4.4.2). I derive this ratio below and indicate its use in a Metropolis-Hastings algorithm to simulate from the posterior of q_k. Expanding the posterior of q_k as shown in equation (5.9) gives

$$p(q_k \mid \ldots) \propto p(q_k) \prod_{vsr: X_{k,vsr} \neq 0} \left\{ \prod_{st: z_{sr}=1} \left[(q_k)^{y_{vsr}} (1 - q_k)^{1-y_{vsr}} \right] \prod_{st: z_{sr}=0} [(1 - y_{vsr})] \right\} \tag{5.18}$$

Consider two values of q_k, say q_k and q'_k. Then the posterior ratio for q'_k compared to q_k is

$$\frac{p(q'_k \mid \ldots)}{p(q_k \mid \ldots)} = \frac{p(q'_k)}{p(q_k)} \prod_{vsr:X_{k,vsr}\neq 0} \prod_{sr:z_{sr}=1} \frac{(q'_k)^{y_{vsr}}(1-q'_k)^{1-y_{vsr}}}{(q_k)^{y_{vsr}}(1-q_k)^{1-y_{vsr}}} \tag{5.19}$$

The Metropolis-Hastings sampler requires the ratio of posteriors. It also requires a proposal distribution (independent of this posterior ratio) for generating new q'_k given the current q_k. (See Section 4.4.2 for details.) Suitable proposal distributions for q_k are defined on the unit interval. These include a uniform distribution either wrapped (rotated) or truncated to the unit interval, or a Gaussian distribution truncated to the unit interval. See Section 4.4.2 for a catalogue of the proposal ratios for these distributions. Recall that the acceptance probability for determining whether to accept the new q_k or not is defined as

$$\mathcal{A}(q'_k \mid q_k) = \frac{p(q'_k \mid \ldots)}{p(q_k \mid \ldots)} \frac{r(q'_k \mid q_k)}{r(q_k \mid q'_k)}. \tag{5.20}$$

For the rotated uniform, the proposal ratio evaluates to unity, which simplifies computations.

The Metropolis-Hastings sampling algorithm for q_k is

Step 1 Initialize q_k to $q_k^{(0)}$.

Step 2 Repeat the following steps for iterations $n = 1, \ldots, N$

Step 2a Generate new proposed value q'_k from the proposal distribution $q'_k \sim R(q'_k \mid q_k^{(n-1)})$.

Step 2b Evaluate $\mathcal{A}(q'_k \mid q_k)$ as per equation (5.20) using the most up-to-date values $z^{(n-1)}$ for v, s, r not updated yet and $z^{(n)}$ for v, s, r already updated.

Step 2c Generate uniform variate $U \sim \text{Uniform}(0,1)$. If $U \leq \mathcal{A}(q'_k \mid q_k)$ then update with new proposed value $q_k^{(n)} = q'_k$ else retain old value $q_k^{(n)} = q_k^{(n-1)}$.

Step 3 Check convergence and efficiency of algorithm (referring to Section 4.4.7). If necessary, repeat Step 2 until convergence to equilibrium is achieved with sufficient samples.

5.5 Pilot simulation experiment

Recall that the three-tier model of this chapter conditions on a fixed value of θ. Prior knowledge of the spatio-temporal presence on the lattice was vague for the *dingo* case study, so biological assessment of results should take this into account. This pilot study provides preliminary results to assist in the implementation of the larger study of Chapter 7. Besides assessing the overall performance of the MCMC simulation procedure, a sensitivity analysis compared performance for different values of parameters in the prior distribution for dingo presence.

Chain design issues, such as starting values used for initializing Markov chain Monte Carlo simulations from equation (5.1), are addressed in Section 5.5.1. Statistics from posterior distributions that are used for monitoring performance of the chains are also defined in this section. Chain calibration issues, such as the length of the chain, burnin, and thinning, are discussed in this section, and a specific experiment to check assumptions is presented in Section 5.5.5. Preliminary results for inference are summarized in Section 5.5.2. Another experiment was specifically aimed at sensitivity analysis of the effect of the proposal distribution $R(q'_k \mid q_k)$ in Section 5.6.

5.5.1 Initialization

It was necessary to choose appropriate values of simulation length T, initial transient T_0, starting values for parameters $q^{(0)}$ and $z^{(0)}$.

Initially a simple form of the proposal distribution $R_1(q_k' \mid q_k)$ was selected: a rotated uniform distribution, with a half-width $h = 0.10$. Initially, a pilot run of length $T = 2,000$ was performed. Another longer run (10,000) was investigated later to see if the equilibrium distribution was affected. Initially, no transient was discarded $T_0 = 0$, in order to determine the appropriate burnin period for further runs. A burnin of 200 was settled on. Starting values $\{q_k^{(0)}\}$ were set to what were considered high values of 0.5. Starting values $\{z_{sr}^{(0)}\}$ were chosen randomly from $0, 1$.

The parameter θ may be selected (following Section 5.3.1) by specifying the three quantities $\hat{p}_{0,0}$, $\hat{p}_{2,2}$, and ρ. The quantity $\hat{p}_{0,0}$ is the estimated overall probability that a dingo visits a position given that there were no visits to neighbouring positions. Values of 0.10 and 0.15 were chosen in line with biological experts' opinions discussed in Section 2.3. The quantity $\hat{p}_{2,2}$ is the estimated overall probability that a dingo visits a position given that there were visits to all neighbouring positions. The value 0.75 was considered conservative yet representative. Values of the ratio of spatial to temporal dependence considered were $\rho = 4, 2, 1, 0.5$.

Sets of dingo prior parameters $\theta_{(m)}$ corresponding to these quantities were calculated using results from Equations 5.11–5.14 and are shown in Table 5.1.

Table 5.1: Choice of parameters in the prior distribution of dingo presence.

Index m of $\theta_{(m)}$	Specified probabilities			Dingo prior parameters		
	$\hat{p}_{0,0}$	$\hat{p}_{2,2}$	$\rho = \frac{\theta_1}{\theta_2}$	$\theta_{(m)0}$	$\theta_{(m)1}$	$\theta_{(m)2}$
1	0.10	0.75	4	-2.1972	1.3183	0.3296
2	0.10	0.75	2	-2.1972	1.0986	0.5493
3	0.10	0.75	1	-2.1972	0.8240	0.8240
4	0.10	0.75	0.5	-2.1972	0.5493	1.0986
5	0.15	0.75	4	-1.7346	1.1333	0.2833
6	0.15	0.75	2	-1.7346	0.9444	0.4722
7	0.15	0.75	1	-1.7346	0.7083	0.7083
8	0.15	0.75	0.5	-1.7346	0.4722	0.9444

Convergence to the posterior (equilibrium) distributions of parameters is ascertained by monitoring parameters and summary statistics from these distributions.

Each of the MCMC chains for $q_k, k = 1, \ldots, 6$ was monitored for convergence properties. However, it would be infeasible to monitor all of the MCMC chains for z_{sr} since these numbered nearly 850 unknown components altogether. Thus only 15 spatio-temporal positions were selected. These were located (in spatio-temporal geometry) in 5 groups with varying amounts of observed dingo visits in neighbouring positions, ranging from no dingos in the vicinity to being surrounded by visits. Figure 3.1 shows the pattern of observed dingo visits which was used to select these positions, and Table 5.2 identifies the positions at which presence was to be sampled.

Also monitored was the dingo presence summed over both spatial and temporal first

order neighbours for each of these selected positions. This is the $h_{sr}(z_{sr}, z_{-sr}, \theta)$ statistic defined in equation (5.6) which is a major contributor to the posterior distribution, the Gibbs sampler and therefore to the simulated values in individual chains $\{z_{sr}^{(t)}\}$. According to Gelman & Meng (1996) canonical statistics such as this should be monitored due to their impact on samplers.

Table 5.2: Spatio-temporal positions selected for monitoring dingo presence.

		position		observed dingo visits to neighbouring positions				No. of nbrs	
		site	day					spatial	temporal
grp	j	s_j	r_j	y_{s_j-1,r_j}	y_{s_j+1,r_j}	y_{s_j,r_j-1}	y_{s_j,r_j+1}	s	r
1	1	11	1	1	1	N/A	0	2	0
	2	11	2	0	0	0	0	0	0
	3	12	2	0	1	1	0	1	1
2	4	25	4	0	1	0	0	1	0
	5	26	5	0	0	1	0	0	1
	6	27	4	1	1	0	0	2	0
3	7	48	5	0	0	1	0	0	1
	8	48	6	0	1	0	0	1	0
	9	49	5	0	0	1	1	0	2
	10	50	5	0	1	0	1	1	1
4	11	69	5	0	0	0	0	0	0
5	12	112	3	1	0	1	0	1	1
	13	112	4	1	0	0	0	1	0
	14	113	3	0	0	0	0	0	0
	15	113	4	0	1	0	0	1	0

5.5.2 Results from pilot experiment

The aim of the inference and computation has been to approximate the posterior distribution of parameters. First MCMC diagnostics are presented to confirm that the chains have converged to the target posterior distributions. Then the posterior distributions are presented and discussed below. In addition, a sensitivity analysis to the choice of $\theta_{(m)}$ is given.

5.5.3 Posterior density of q_k

Figure 5.2 is a simple time-series of the values of $q_1^{(t)}$ obtained at each MCMC iteration. There seems to be a fairly high correlation between adjacent values (as evidenced by the high IACT values), although the overall mean appears fairly stationary. The posterior density of q_1 appears to be fairly Gaussian, centred around its mean 0.54, with a standard deviation of 0.07, and most values lying between 0.40 and 0.70.

The time-series of the MCMC estimates of $q_k, k = 1, \ldots, 6$ depicted in Figure 5.3 show similar behaviour. In particular, $q_2^{(t)}$ has a smaller variance than the others, but also has a lower mean value.

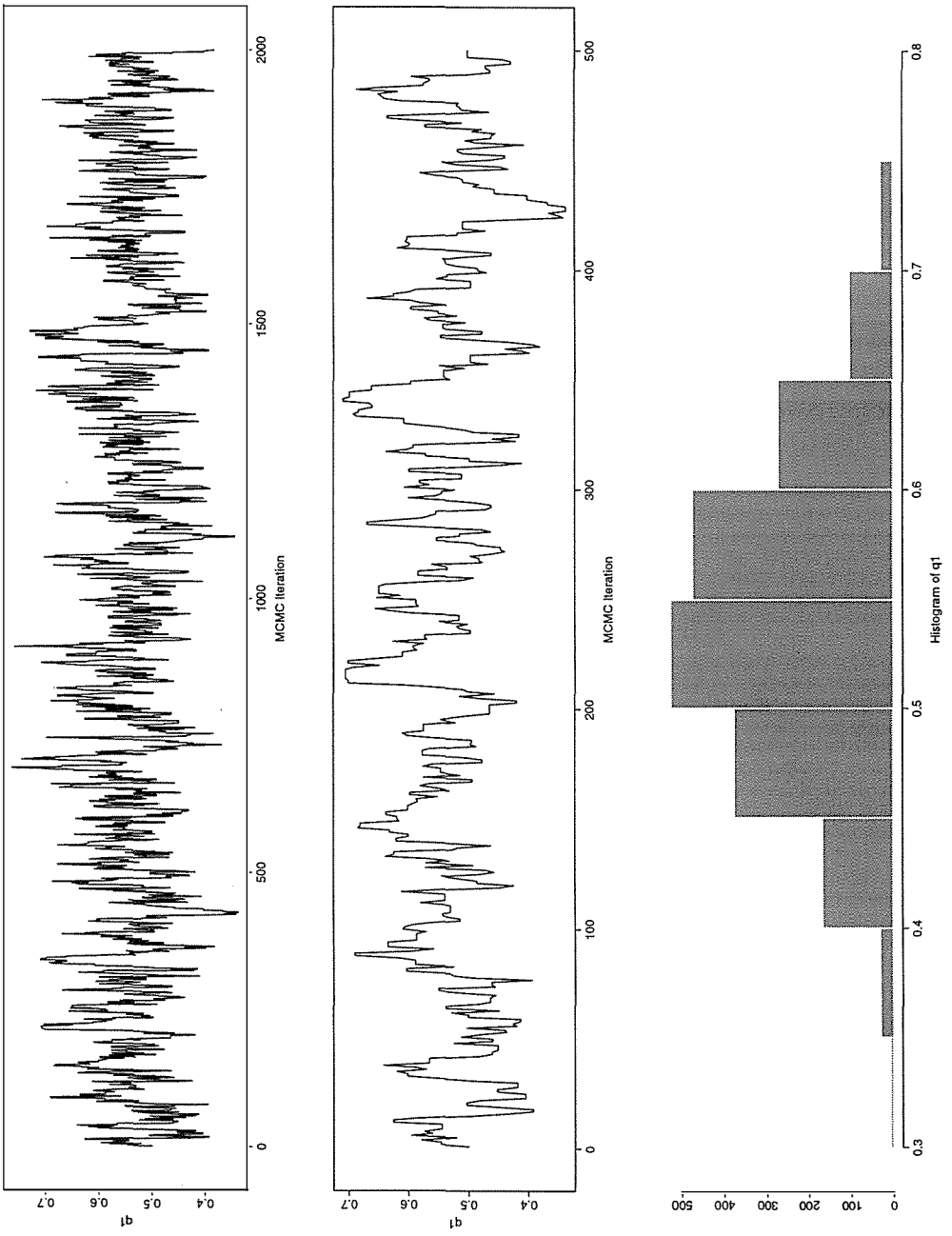


Figure 5.2: Pilot run: Posterior distribution of q_1. From top to bottom, plots are: (a) time-series of $q_1^{(t)}$ plotted against MCMC time $t = 1, \ldots, T$; (b) close-up of first 500 iterations; (c) histogram of posterior distribution of q_1.

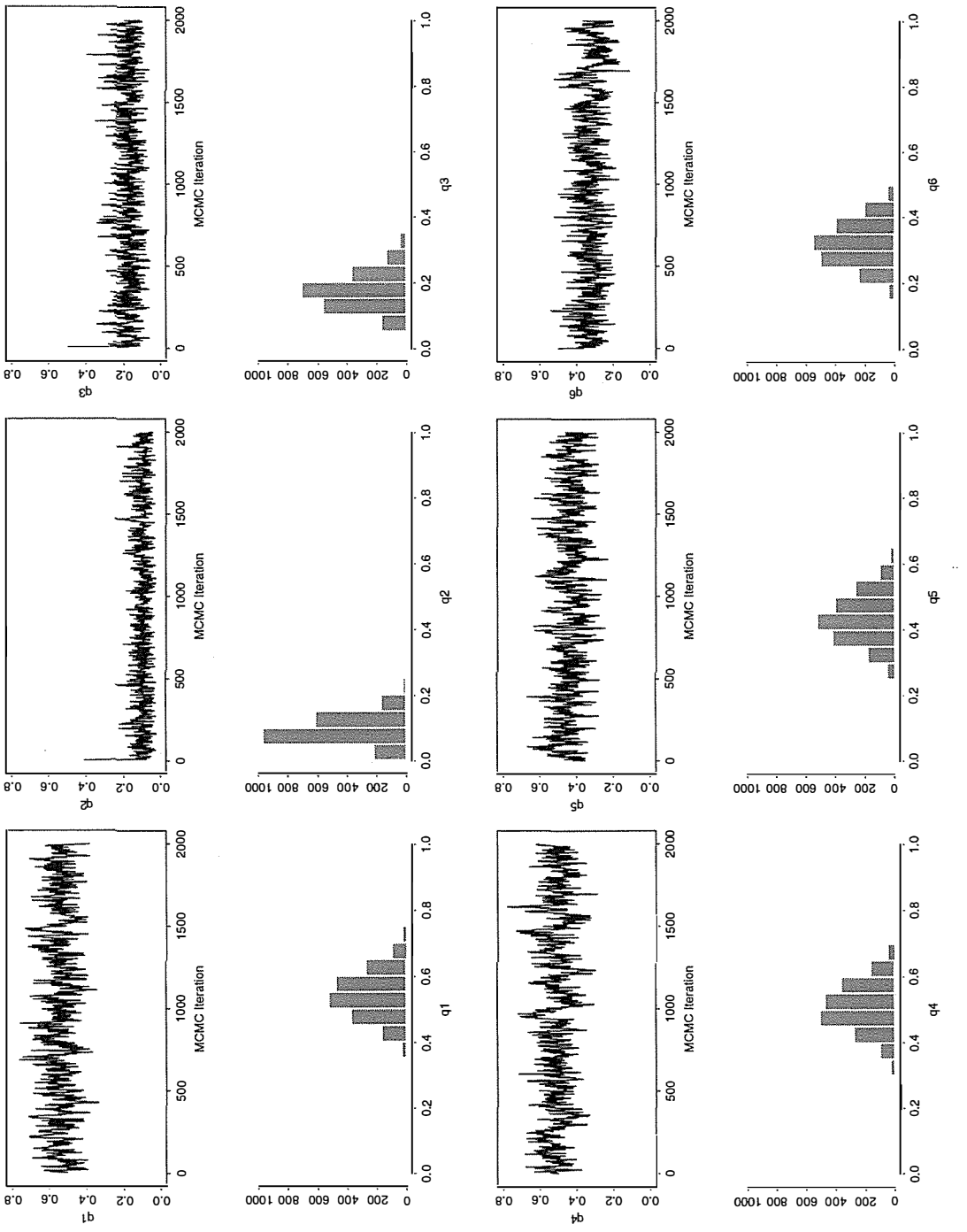


Figure 5.3: Pilot run: Posterior distribution of $q_k, k = 1, \ldots, 6$. In pairs of (a) time-series of $q_k^{(t)}$ plotted against MCMC time $t = 1, \ldots, T$ and (b) histogram of posterior distribution of q_k. From left to right, top to bottom, pairs of plots for $q_k, k = 1, \ldots, 6$.

The chains for each q_k were divided into batches of 100, and batch means and variances investigated. Plots of these for parameter set 1 are shown in Figures 5.4 and 5.5. The plots indicate that the first batch of size 100 should be discarded as the means and variances are 'significantly' different than those for other batches. In general, the behaviour of q_2 and q_3 was more consistent over batches, with q_1, q_4, q_6 being most inconsistent. The first set of plots support results from the Geweke and Hiedelberger & Welsh tests for stationarity, which are not shown here. Note that no standard MCMC tests have been devised for testing stationarity in variance across batches, although Geweke's test could easily be extended to achieve that.

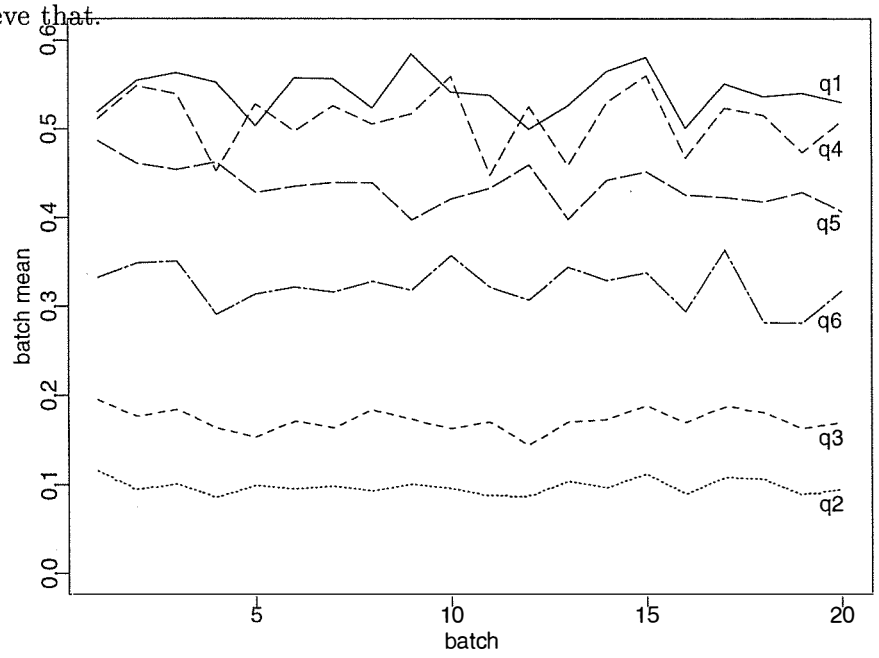


Figure 5.4: Pilot run: batch means of q_k. Batch size 100.

The integrated autocorrelation times range from 0.23%–0.85% of the total number of iterations, T. These all lie below the recommended 1% of T (Green & Han 1990), suggesting that 2000 runs are sufficient to ensure that we are using essentially independent samples for inference. This was confirmed by exploratory experiments with more iterations (Section 5.5.5).

The acceptance rates for q_k are quite high, ranging from 53.2% for the narrowest posterior distribution, to 72%, indicating an adequate choice of the proposal distribution.

Statistics from the posterior density of q_k were calculated for each of the eight parameter sets described in Table 5.1. These included the mean, standard deviation, integrated autocorrelation time (IACT) and associated minimum bandwidth, and % acceptances in the Metropolis-Hastings updates. Representative results are tabulated below in Tables 5.3 for just one selection $\theta_{(1)} = [-2.1972, 1.3183, 0.3296]$. Results obtained from other parameters $\theta_{(m)}$, $m = 1, 2, 3, 4$ were similar so are not tabulated.

As to be expected from the preliminary analysis, Chemical A had the highest attractiveness ability, and Chemical B the lowest. However, Chemical D was a very close contender for

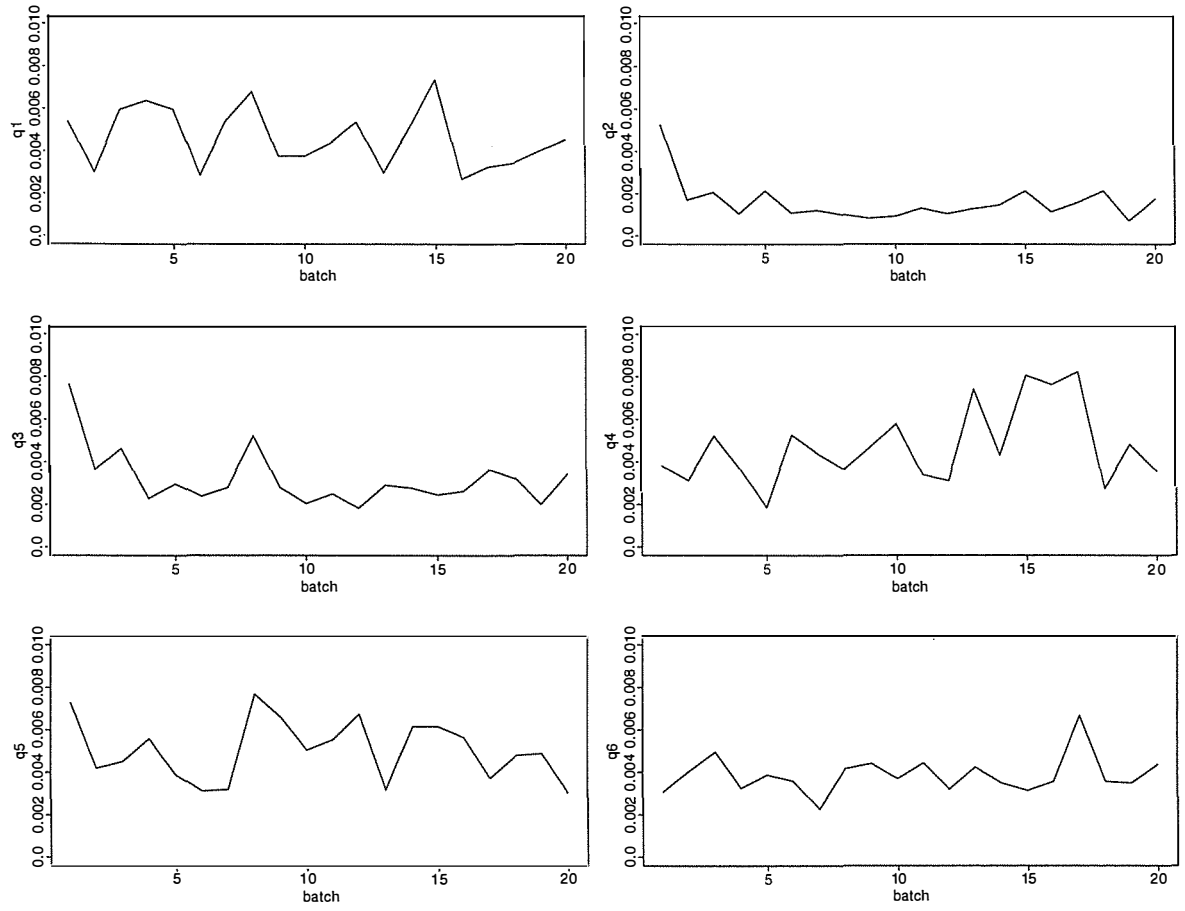


Figure 5.5: Pilot run: batch variances of q_k. Batch size 100.

Table 5.3: Pilot run: Statistics from posterior density of q_k.

Chemical	Statistics from posterior density of q_k		
k	mean (stdev)	IACT (M)	%acc
A	0.5410 (0.07180)	11.50 (35)	69.5
B	0.0973 (0.04056)	4.69 (15)	53.2
C	0.1724 (0.05757)	5.86 (18)	62.2
D	0.5094 (0.07620)	17.70 (54)	70.0
E	0.4352 (0.07413)	12.17 (37)	71.8
F	0.3227 (0.06620)	9.87 (30)	68.4

first place, which was not so evident from preliminary findings (Chapter 3). It was followed (not so closely) by Chemical E and then Chemical F, which had reasonable attractiveness ability. Chemical C was second-last on this ordering.

A graphical method often used in multivariate problems, the starplot, will be used to compare the relative sizes of groups of estimates, for instance $\{q_k\}$, between parameter sets. Figure 5.6 illustrates the differences in median estimates of q_k between parameter sets. The length of each radius in each starplot represents the mean value of the posterior density of one of the variables. Each starplot then shows the relationship between the variables in a particular set of dingo prior parameters. Since all the starplots have virtually the same shape, this indicates that the relationship between values of $q_k^{(t)}$ was virtually the same for each parameter set $\theta_{(m)}$. Thus the relative sizes of median posterior estimates of $q_k^{(t)}$ are fairly robust to changes in the parameters $\theta_{(m)}$ in the dingo density prior.

In contrast, the absolute sizes of the medians of the posterior distributions of q_k vary according to parameter $\theta_{(m)}$. Thus prior expectations on the prevalence and patterns of underlying dingo presence have a large impact on the absolute power of attraction for the chemicals. It is clear that estimates based on $\theta_{(1)}, \ldots, \theta_{(4)}$ (with $\widehat{p}_{0,0} = 0.10$) are about 30% larger than those based on $\theta_{(5)}, \ldots, \theta_{(8)}$ (with $\widehat{p}_{0,0} = 0.15$).

Thus in the first case increased frequency of dingo visits is attributed by the model evenly across all chemicals, whereas in the second case it is attributed to underlying increase of dingo presence (15%) regardless of chemicals present. As mentioned in Weir & Pettitt (1999) this exemplifies a common dilemma with spatial modelling where spatial coefficients and covariate coefficients compete.

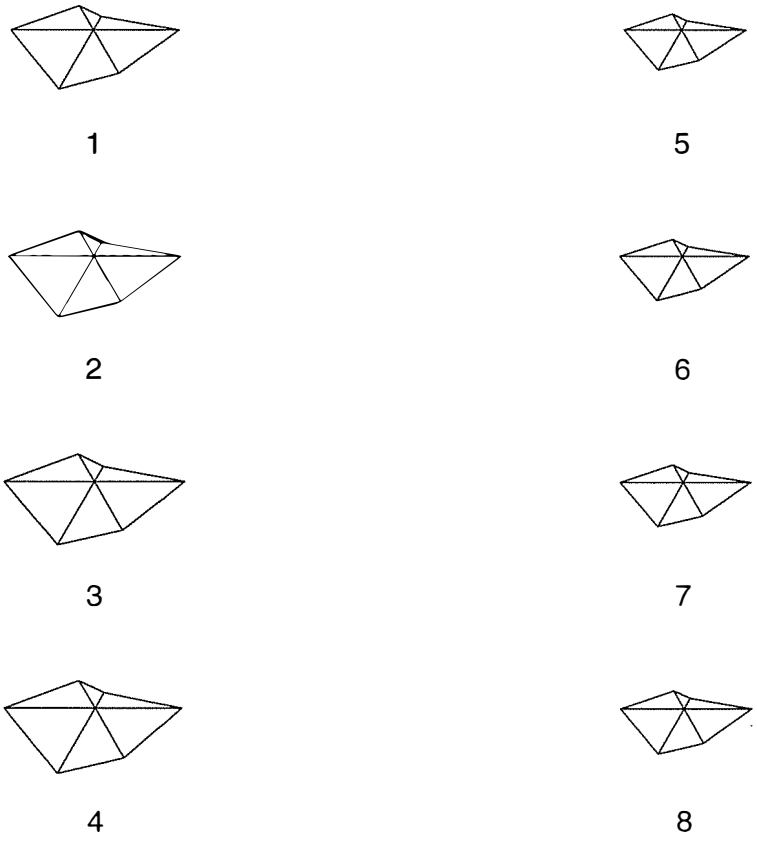


Figure 5.6: Pilot run: Starplots showing comparison of parameter estimates $\hat{q}_k$ between runs with different parameters in prior for dingo presence $\theta_{(m)}$. Numbers refer to index m. Each radius of a star corresponds to one of the q_k parameters, starting with q_1 at the eastward pointing radius, and cycling anti-clockwise through $k = 2, 3, \ldots, 6$.

5.5.4 Posterior density of z_{sr}

We present results here for a particular value of the prior parameter $\theta_{(1)}$, although results are similar across the range of $\theta_{(m)}$ investigated. The mean posterior density of dingo presence over the entire spatial-temporal lattice was 21.69%.

Table 5.4: Pilot Run: Statistics from posterior density at selected positions of dingo presence $z_{s_j r_j}$ and summed over neighbouring positions, as described in Table 5.2. Results shown are representative for different $\theta_{(m)}$ and correspond to $\theta_{(1)}$.

			Statistics from posterior density of $z_{s_i r_i}$					Statistics from posterior density of $h_{sr}(z_{s_i r_i}, z_{-s_i r_i}; \theta)$			
	nbrs										
pos	h	v	mean	stdev	IACT	M	%acc	mean	stdev	IACT	M
1	2	0	0.5645	0.2460	0.94	3	50.2	0.7108	0.0128	1.27	4
2	0	0	0.1325	0.1150	1.27	4	20.6	0.2353	0.0366	1.15	4
3	1	1	0.2170	0.1700	1.19	4	31.3	0.5496	0.0116	1.30	4
4	1	0	0.2385	0.1817	1.20	4	34.5	0.3640	0.0249	1.19	4
5	0	1	0.1110	0.0987	1.12	4	19.5	0.5375	0.0087	1.18	4
6	2	0	0.3025	0.2111	1.04	4	42.3	0.5425	0.0094	1.05	4
7	0	1	0.0900	0.0819	1.31	4	14.3	0.3521	0.0226	1.41	5
8	1	0	0.1625	0.1362	1.22	4	25.3	0.3111	0.0150	1.22	4
9	0	2	0.1660	0.1385	1.34	5	23.3	0.6150	0.0208	1.26	4
10	1	1	0.3700	0.2332	1.04	4	43.0	0.6151	0.0224	1.32	4
11	0	0	0.0650	0.0608	1.30	4	11.2	0.0610	0.0146	1.21	4
12	1	1	0.2970	0.2089	1.17	4	39.6	0.6024	0.0197	1.32	4
13	1	0	0.2935	0.2075	1.17	4	37.2	0.4463	0.0378	1.41	5
14	0	0	0.1160	0.1026	1.17	4	19.0	0.2478	0.0505	1.21	4
15	1	0	0.2470	0.1861	1.11	4	32.9	0.3833	0.0273	1.36	5

As shown in Table 5.4, the mean of the posterior density of a given position appears to increase with increasing number of spatial neighbours h, yet is not much affected by the number of temporal neighbours v, indicating that perhaps spatial dependence is more important than temporal. Note that these statistics do not consider the effect of the neighbourhood beyond first order neighbours.

The maximum IACT at any position is 1.34, or .067% of the number of iterations, indicating that the chain comprises sufficiently independent observations. This statistic needs to be interpreted with care in this context since z_{sr} can only adopt 1 of 2 values. The acceptance rates tend to be proportional to the estimated probability of dingo presence, ranging from 50% for position 1 (2 spatial neighbours) with the highest probability of dingo presence, to 11% for isolated position 11 (no neighbours at all) with the lowest. These statistics are supported by plots of the time series trace of dingo presence, not shown here because the summary statistics reflect their contents.

Figure 5.7 gives an overall picture of the estimated dingo presence over the entire area, and uses greyscale to show the varying sizes of these estimates. Dingo presence is estimated to be larger if the position is situated close to positions where visits were actually observed, the more such positions, the higher the estimated probability. Notice how dingo presence

tends to "spread" horizontally across the area from positions of observed visits. The horizontal spread is due to the positive value of θ_1, spatial dependence, in the prior distribution. Vertical spread over all days appears to often apply where a dingo visited that site on any single occasion.

Corresponding Figure 5.8 for dingo presence averaged over neighbouring positions shows a smoother version of this. Investigation of the estimated neighbourhood averages shows potential impact of neighbourhood on estimated probability of dingo presence. This diagram shares some of the general spatio-temporal trends of Figure 5.7. One discrepancy is that posterior probabilities are more polarised which is to be expected when comparing point probability to average neighbour probability. However temporal striations and isolated incidences of higher posterior probabilities provide evidence that chemcials are providing some information to the analysis.

Comparisons between parameter sets, Figure 5.10 and 5.11, show that the amount of spatial dependence in the model definitely affects the relative sizes of the estimates of dingo presence at these selected positions. As the eye travels down the figure, there is less spatial dependence and more temporal dependence allowed in the prior for dingo presence. (That is, θ_1 decreases relative to θ_2 going down the page.) The positions where there is a higher number of spatial neighbours as identified by the spikes in Figure 5.9(b) are most prominent for larger θ_1. Those positions where there is a higher number of temporal neighbours as identified by the spikes in Figure 5.9(c) are most prominent for larger θ_2.

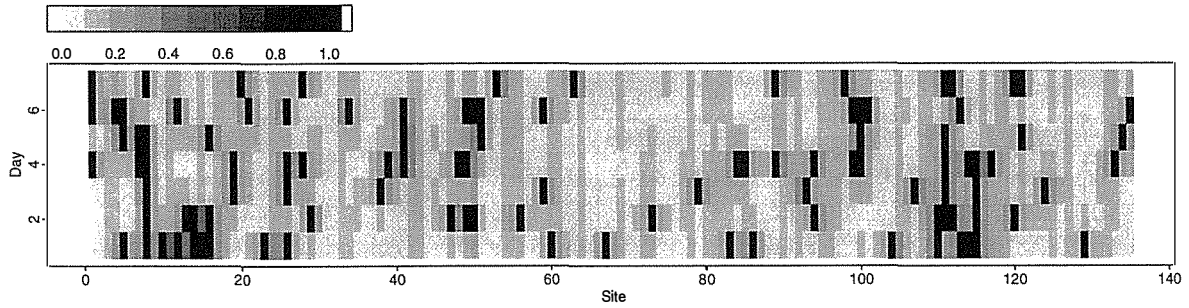


Figure 5.7: Pilot run: posterior probability of a dingo being present. Black represents occasions where dingo visits were actually observed, so were definitely present. Grey sites correspond to occasions where dingo visits were not observed, so dingo presence had to be estimated. Shades of grey represent the posterior probability of dingo being present.

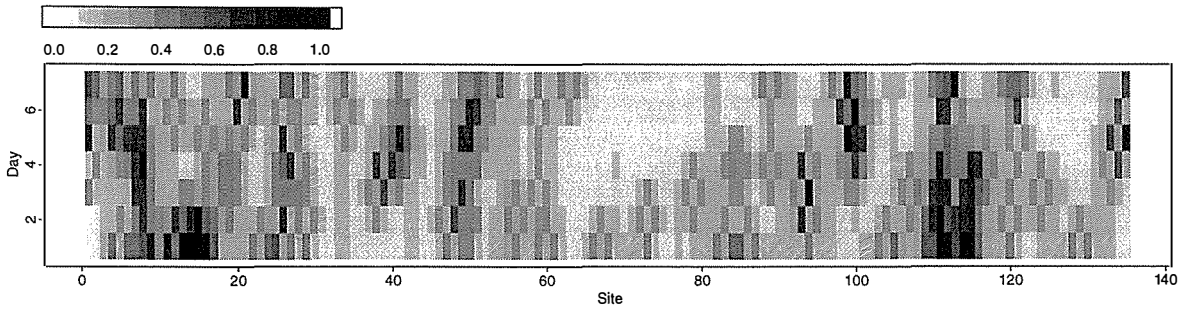


Figure 5.8: Pilot run: posterior probability of a dingo being present, averaged over first order neighbouring positions. Black represents high probability of dingos being present in neighbouring positions and lighter grey lower probabilities.

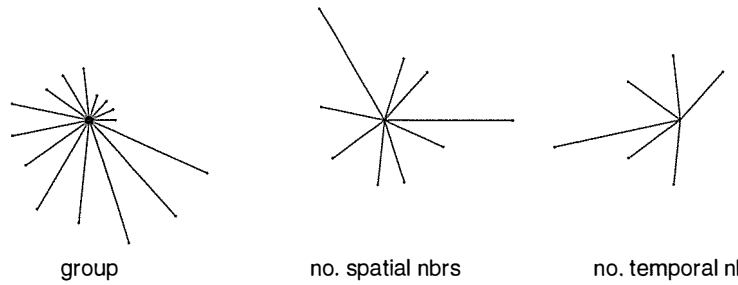


Figure 5.9: Reference to features of positions on the lattice monitored for convergence of $z_{sr}^{(t)}$. Positions were located in (a) groups; with (b) various numbers of spatial neighbours h, and (c) various numbers of temporal neighbours v. To be used in conjunction with starplots in figure 5.10.

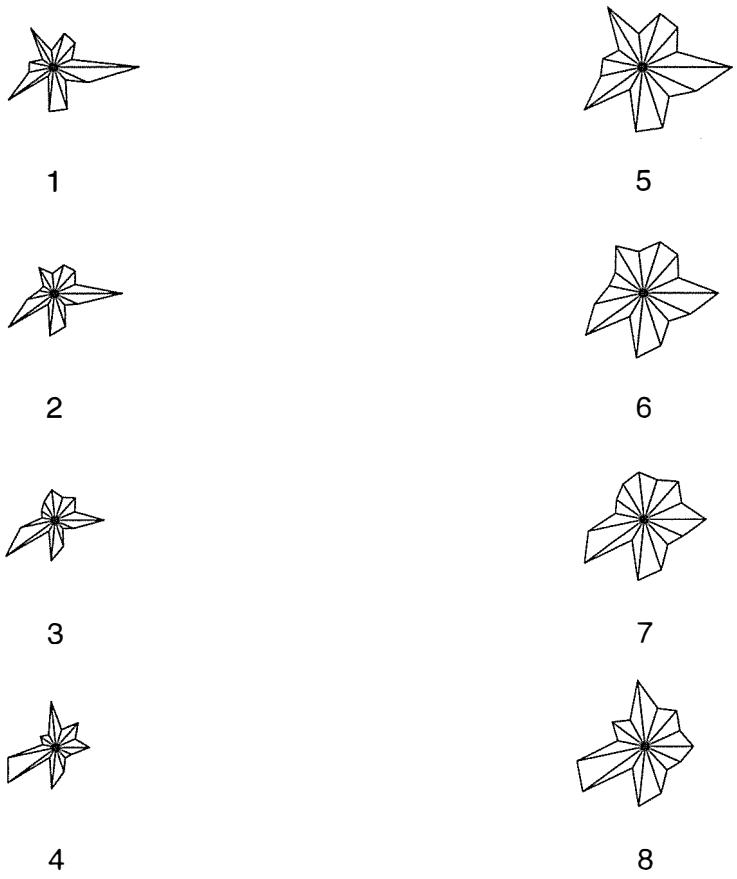


Figure 5.10: Pilot run: Starplots showing comparison of parameter estimates $\hat{z}_{s_j r_j}$, between runs with different parameters in the prior for dingo presence $\theta_{(m)}$. Numbers below plots refer to index m. The eastward pointing radius of each starplot represents $\hat{z}_{s_1 r_1}$, and cycles anti-clockwise through $j = 2, 3, \ldots$.

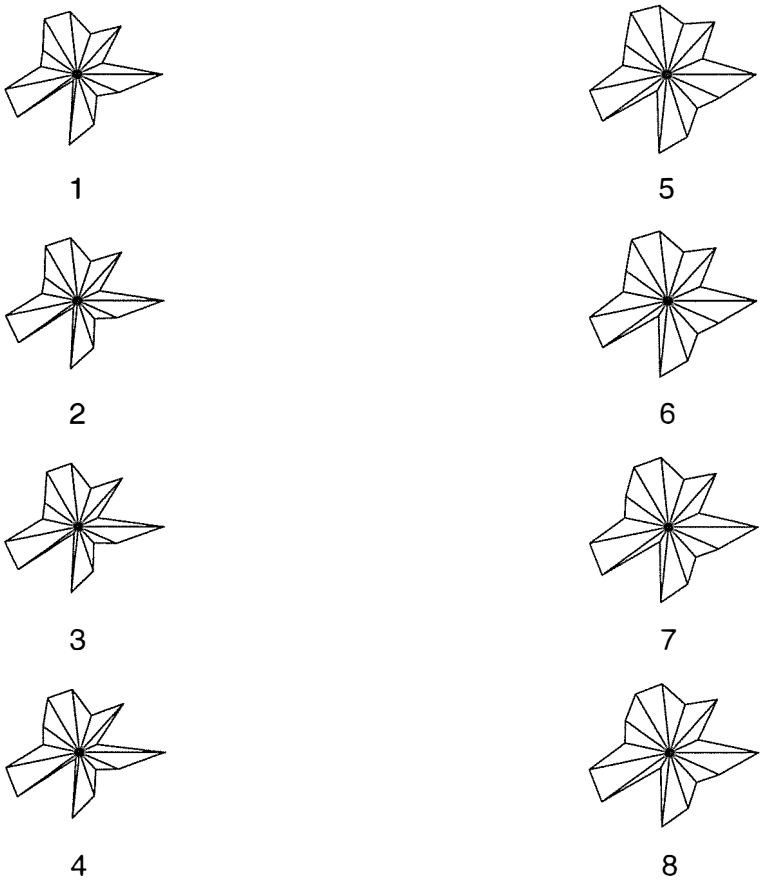


Figure 5.11: Pilot run: Starplot showing comparison of parameter estimates $\hat{z}_{s_j r_j}$, averaged over neighbouring positions, between runs with different parameters in the prior for dingo presence $\theta_{(m)}$. Numbers refer to index m. The eastward pointing radius of each starplot refers to the case with $j = 1$, and cycles through $j = 2, 3, \ldots$ in an anti-clockwise direction

5.5.5 Checking convergence

A related experiment with a longer run of length 10,000 iterations was performed. Model parameters were otherwise unchanged from the pilot simulation experiment setup described in Section 5.5.1. Sensitivity analysis with respect to parameters $\theta_{(m)}$ covered $m = 1, 2, 3, 4$. Different starting values for the dingo presence parameters $\{z_{sr}^{(0)}\}$ were considered:

- *hot* values (all sites indicating dingo presence);
- *cold* values (no unknown sites indicating dingo presence);
- *sample* presence from the posterior distribution $p(z_{sr} \mid \ldots)$ from previous experiment;
- *random* values, with even odds of dingo being present or not.

From the results shown in Tables 5.5, 5.6 and 5.7, we can see that the estimates and acceptance rates of $\hat{q}_k$ differ by less than 2% using different starting values for dingo presence parameters. For random sampling values the minimum, median, and maximum IACT for q_k was consistently lowest given parameter for dingo presence prior $\theta_{(1)}$. Thus an improvement in convergence, without a change in estimates and acceptance rates, means that the random sampling approach is the best option.

Tables 5.8 and 5.9 showed that for different parameter sets the maximum and median IACT of d_{it} calculated over selected dingo positions was most often lowest when 'hot' starting values were used for dingo presence. However, the convergence time for these parameters shows substantial improvement, so the convergence time of $\{q_k\}$ parameters should determine the best starting values for dingo presence.

These results suggested that random starting values for $\{z_{sr}\}$ should be used, and that a run length of 10,000 simulations was adequate.

Table 5.5: Posterior mean of q_k for different starting values in dingo presence $\{z_{sr}^{(0)}\}$

$\{z_{sr}^{(0)}\}$	$\hat{q}_1$	$\hat{q}_2$	$\hat{q}_3$	$\hat{q}_4$	$\hat{q}_5$	$\hat{q}_6$
hot	0.538006	0.096132	0.171752	0.515784	0.427271	0.324981
cold	0.538668	0.096458	0.172306	0.514164	0.427322	0.318923
sample	0.543666	0.096007	0.173125	0.522828	0.433221	0.323405
random	0.538566	0.094704	0.173314	0.518474	0.428769	0.320972

Table 5.6: Percentage acceptance rates of q_k for different starting values in dingo presence $\{z_{sr}^{(0)}\}$

$\{z_{sr}^{(0)}\}$	q_1	q_2	q_3	q_4	q_5	q_6
hot	67.8	56.6	62.9	69.4	69.3	68.7
cold	67.0	56.8	62.8	70.1	70.0	68.2
sample	67.5	56.8	63.7	70.6	70.2	68.3
random	68.2	57.1	62.9	69.5	70.7	67.9

Table 5.7: IACT: Minimum, median, and maximum over q_k for different starting values in dingo presence $\{z_{sr}^{(0)}\}$. Results are given for $\theta_{(1)}$, which are indicative of other $\theta_{(m)}$.

$\{z_{sr}^{(0)}\}$	minimum	median	maximum
hot	6.815060	11.31607	15.10838
cold	5.562789	11.23415	16.91372
sample	4.855034	11.90683	14.64922
random	4.846942	10.34093	12.95322

Table 5.8: Median IACT over selected dingo positions for different starting values in dingo presence $\{z_{sr}^{(0)}\}$ with different $\theta_{(m)}$.

$\{z_{sr}^{(0)}\}$	$\theta_{(1)}$	$\theta_{(2)}$	$\theta_{(3)}$	$\theta_{(4)}$
hot	1.091483	1.088368	1.118390	1.115506
cold	1.128104	1.147700	1.137382	1.102967
sample	1.141530	1.125565	1.141240	1.122637
random	1.132488	1.133810	1.115549	1.134255

Table 5.9: Maximum IACT over selected dingo positions for different starting values in dingo presence $\{z_{sr}^{(0)}\}$ with different $\theta_{(m)}$.

$\{z_{sr}^{(0)}\}$	$\theta_{(1)}$	$\theta_{(2)}$	$\theta_{(3)}$	$\theta_{(4)}$
hot	1.195761	1.220022	1.220409	1.182483
cold	1.287259	1.313331	1.262367	1.250697
sample	1.247759	1.257294	1.211549	1.311860
random	1.211182	1.287207	1.267278	1.215944

5.6 Proposal experiment

Initially, three different types of proposal distributions $R(q'_k, q_k)$ were considered for updating the chemical attractiveness effects q_k via the Metropolis-Hastings algorithm, as shown in Table 5.10. Theoretical results were provided in Section 4.4.2.

Table 5.10: Proposal distributions investigated for updates of q_k

Notation	Proposal distribution $R(q'_k \mid q_k)$	Defining parameters
R_1	rotated uniform distribution	half-width h
R_2	truncated uniform distribution	half-width h
R_3	truncated gaussian distribution	standard deviation σ

Several different values of the proposal parameters h, σ were investigated to determine the best value to obtain optimum behaviour of the MCMC chain. The half-widths chosen, and the corresponding theoretical standard deviations are shown in Table 5.11.

Table 5.11: Half-width parameters of uniform proposal distributions and corresponding standard deviations of Guassian proposals investigated for updates of q_k. Half-widths range from 0.01 to 0.10.

Uniform half-width	**theoretical variance**	**standard deviation**
0.01	0.00003	0.00577
0.02	0.00013	0.01155
0.03	0.00030	0.01732
0.04	0.00053	0.02309
0.05	0.00083	0.02887
0.06	0.00120	0.03464
0.07	0.00163	0.04041
0.08	0.00213	0.04619
0.09	0.00270	0.05196

Simulations otherwise calibrated as for the Pilot experiment were repeated for each of these parameter values (half-width or corresponding standard deviation, for each of three proposal distributions and for four different values of the parameter $\theta_{(m)}$ in the prior for dingo presence. These correspond to $m = 1, 2, 3, 4$. Note that $\theta_{(1)}$ represents the highest level of spatial dependence relative to temporal dependence (highest θ_1 compared to θ_2) and $\theta_{(4)}$ represents the opposite case with highest temporal dependence relative to spatial. The case of equal spatial and temporal dependence is represented by $\theta_{(3)}$. Parameter $\theta_{(2)}$ allots spatial and temporal dependence at levels between that of $\theta_{(1)}$ and $\theta_{(3)}$.

The aim was to maximise the number of effectively independent observations, which is equivalent to minimising the integrated autocorrelation function for the parameter estimates. Maximum (over values of k) $\tau(\bar{q}_k)$ were compared to select the optimum proposal distribution. With a run length of 2000, a value of at most $\tau(\bar{q}_k) = 2000/100 = 20$ is desired. Higher values indicate that further runs are necessary.

Another parallel aim was to maximise the efficiency of the MCMC algorithm, which is indicated by maximum acceptance rates. Minimum acceptance rates over all parameter estimates were compared to select the optimum proposal distribution for this criteria. Acceptance rates above 50% indicated that the algorithm was accepting at least half of the values generated from the proposal distribution, which can be considered an adequate level.

Thus we need to balance both these requirements of minimising IACT subject to an upper bound of 10, and maximising the acceptance rates subject to a lower bound of 50%.

Results are shown in Figures 5.12 and 5.13. There is a small amount of variation due to simulation, although replications of this experiment yielded similar results, so we present results of just one replication as representative of the general results.

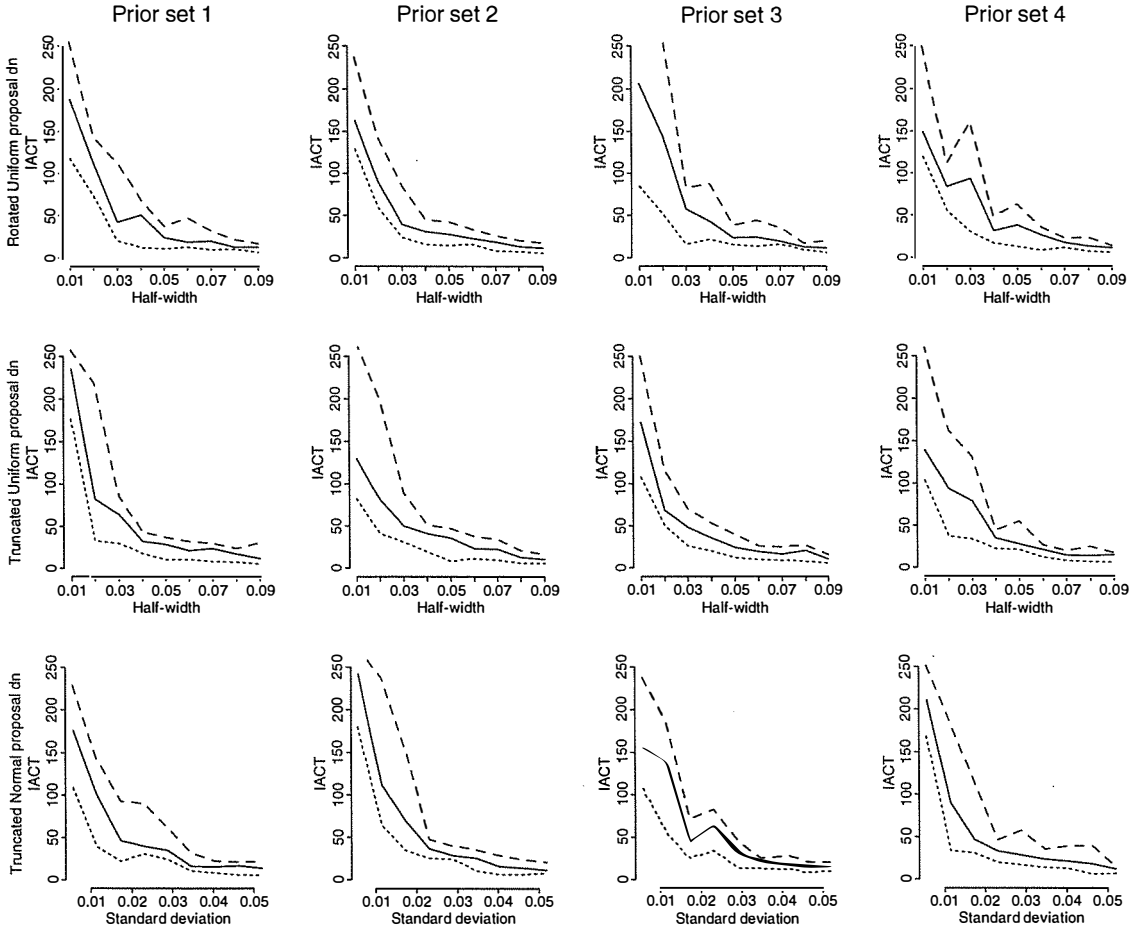


Figure 5.12: IACT of q_k: minimum ($\cdots$), median (—) and maximum (- - -) IACT for various dingo presence prior parameters $\theta_{(m)}$ (columns), and for different proposal distributions $R(q'_k \mid q_k)$ (rows). Each plot depicts change in IACT statistic (y-axis) compared to the characteristic parameter of the proposal distribution R (x-axis). Increasing half-width for uniform proposals correspond roughly to increasing standard deviation of Gaussian proposal.

Results obtained using the truncated uniform appear to be marginally less erratic than those for the rotated uniform, since there was a smaller difference between the maximum and minimum IACT and acceptance rates. In addition, the minimum acceptance rate was usually lower for the rotated than for the truncated uniform proposal, and maximum IACT

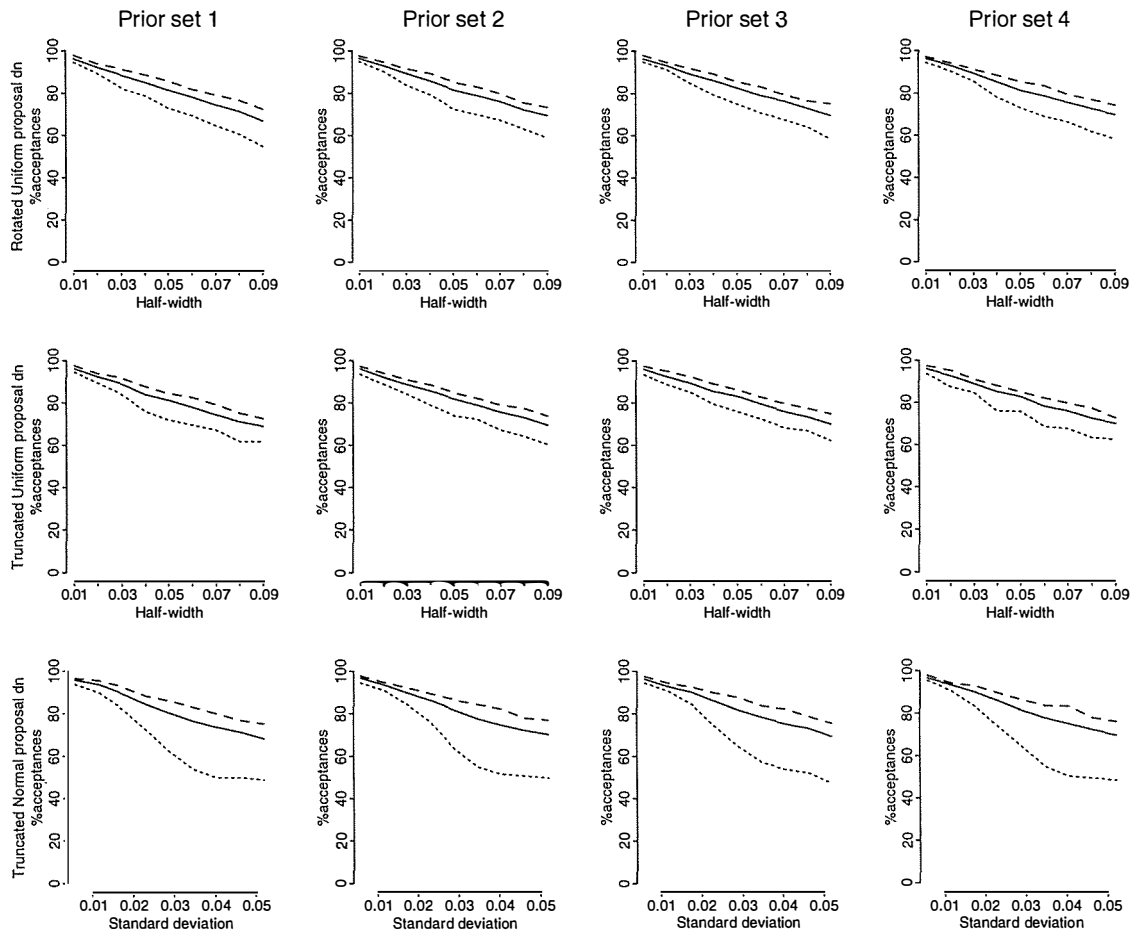


Figure 5.13: %Acceptance rates of q_k: minimum ($\cdots$), median (—) and maximum (- - -) %acceptance rates for various dingo presence prior parameters $\theta_{(m)}$ (columns), and for different proposal distributions $R(q'_k \mid q_k)$ (rows). Each plot depicts change in %acceptance rates (y-axis) compared to the characteristic parameter of the proposal distribution R (x-axis). Increasing half-width for uniform proposals correspond roughly to increasing standard deviation of Gaussian proposal.

was usually higher for the truncated uniform proposal compared to the rotated uniform proposal. These results were consistent over various values $\theta_{(m)}$, $m = 1, ..., 8$ representing different spatio-temporal behaviour. Since the truncated uniform requires a small increase in computing effort, and performs worse on both criteria, the rotated uniform appears to be the better choice.

The difference in results obtained for truncated uniform and truncated Gaussian (with appropriately selected standard deviation), however, do not warrant the larger increase in computing effort. Thus the type of proposal distribution selected is the rotated uniform. It remains to choose the value of the half-width h.

Further experimentation for larger values of the half-width were required, since IACT appeared to be decreasing further and had not yet fallen below 20 and the acceptance rate was still well above 50% for the range of $h = 0.01, 0.02, \ldots, 0.09$.

The results for investigating $h = 0.05, 0.10, \ldots, 0.50$ are presented below in Table 5.12 and Figures 5.14 and 5.15. These figures cover the range $h = 0.01$ to $h = 0.50$, taking into account the change in scale. IACT decreases as it approaches $h = 0.15, 0.20$, and achieves a maximum of 20 for $h = 0.10$. Minimum acceptance rates remain above 50% for $h > 0.10$, and begins to decline rapidly for larger values of h. Thus it appears that the initial choice of $h = 0.10$ is ideal. Higher values of h have poorer acceptance rates, and thus poorer computing efficiency. Lower values of h led to excessively large IACT, requiring longer simulations and hence more computational resources.

Table 5.12: Half-width parameters of uniform proposal distributions and corresponding standard deviations of Guassian proposals investigated for updates of q_k. Half-widths range from 0.05 to 0.50.

Uniform half-width	theoretical variance	standard deviation
0.05	0.00083	0.02887
0.10	0.00333	0.05774
0.15	0.00750	0.08660
0.20	0.01333	0.11547
0.25	0.02083	0.14434
0.30	0.03000	0.17321
0.35	0.04083	0.20207
0.40	0.05333	0.23094
0.45	0.06750	0.25981
0.50	0.08333	0.28868

5.6.1 Conclusion

As a consequence of the above investigations, a truncated uniform proposal distribution with half-width $h = 0.10$ was chosen for the Metropolis-Hastings updates of q_k. It performed marginally better than several choices of θ in the prior for dingo presence. The rotated uniform proposal distribution is computationally simpler and performs only slightly worse, so is also a good candidate. A sample size of 2,000 is sufficient for a particular $\theta_{(m)}$ for estimating the posterior distributions of $\{q_k\}$ or $\{z_{sr}\}$.

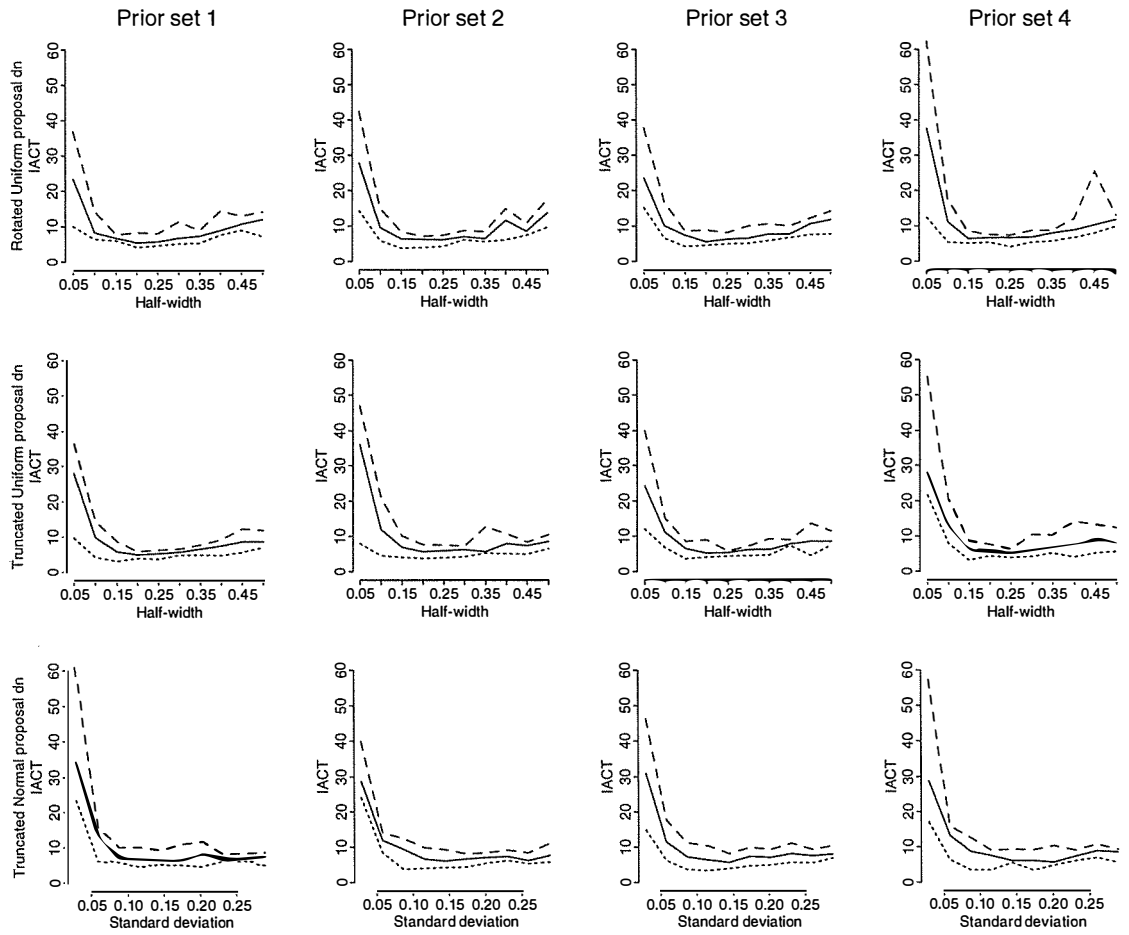


Figure 5.14: IACT of q_k: minimum ($\cdots$), median (—) and maximum (- - -) IACT for various dingo presence prior parameters $\theta_{(m)}$ (columns), and for different proposal distributions $R(q'_k \mid q_k)$ (rows). Each plot depicts change in IACT statistic (y-axis) compared to the characteristic parameter of the proposal distribution R (x-axis). Increasing half-width for uniform proposals correspond roughly to increasing standard deviation of Gaussian proposal. Half-widths in the range 0.05 to 0.50.

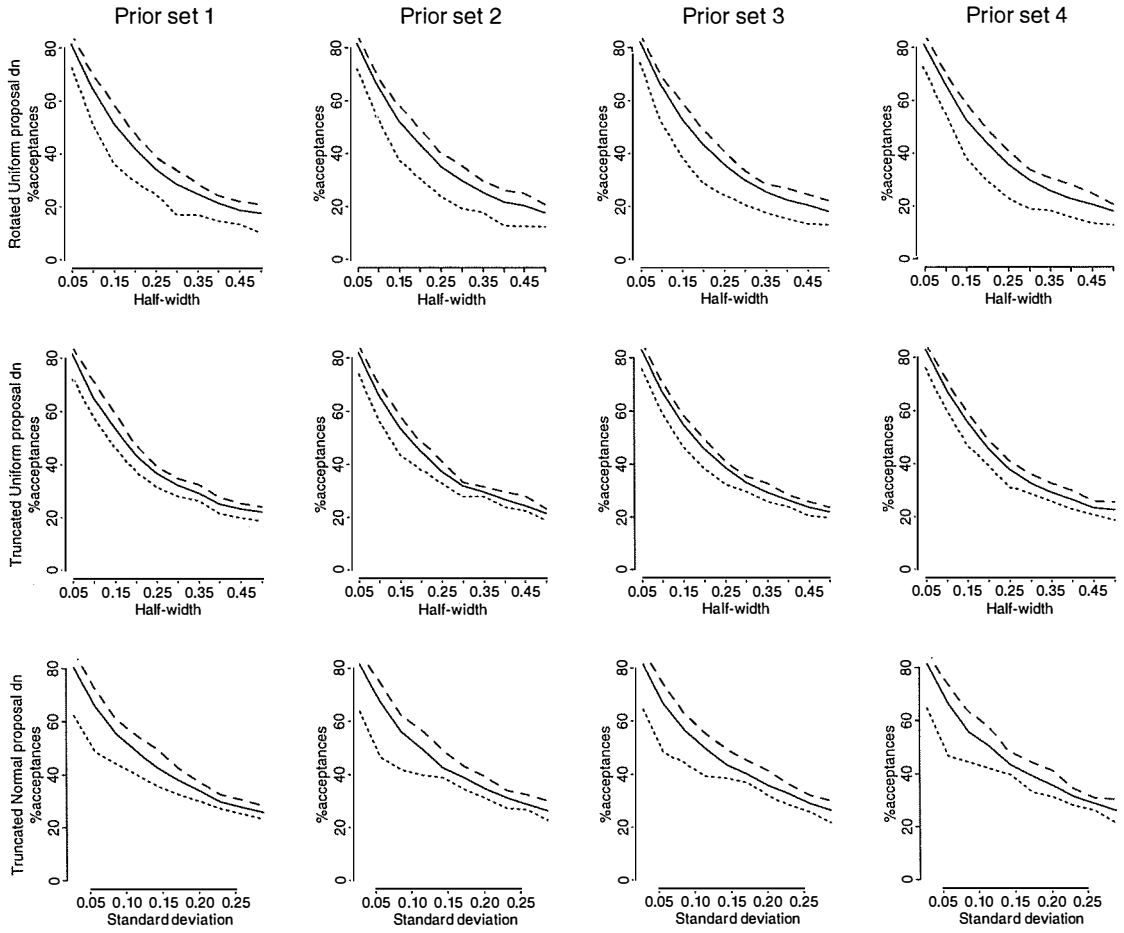


Figure 5.15: %Acceptance rates of q_k: minimum ($\cdots$), median (—) and maximum (- - -) %acceptance rates for various dingo presence prior parameters $\theta_{(m)}$ (columns), and for different proposal distributions $R(q'_k \mid q_k)$ (rows). Each plot depicts change in %acceptance rates (y-axis) compared to the characteristic parameter of the proposal distribution R (x-axis). Increasing half-width for uniform proposals correspond roughly to increasing standard deviation of Gaussian proposal. Half-widths in the range 0.05 to 0.50.

5.7 Bayesian model choice

In the last section, Figures 5.6 and 5.10 showed that the values of the dingo presence prior parameters θ did not appreciably affect the *relationship* between the estimated chemical effects $\{q_k\}$, but did affect the *magnitude* of these effects. In addition, the balance between θ_1 and θ_2 reflects the relative propensity for spatial and temporal correlation in the movement of the dingoes, as illustrated by the biologists' dilemma in Figure 2.2. So the aim in this section is to determine 'good' values of these parameters according to these two criteria: parameter values best supported by the data, thus leading to better estimates of the chemical effects; and parameter values accurately representing the spatio-temporal behaviour of the dingoes.

More generally, an auto-logistic /Ising model, for 2-D data on a regular lattice, was described in Chapter 4. The size of the neighborhood influencing prescence and absence affects the number of parameters included in the model. A major problem in estimation is then two-fold, and can be expressed in a Bayesian fashion:

1. What is the *minimum* number of parameters (i.e. size of neighborhood) which will sufficiently capture the pattern of spatial dependence in this dataset?

2. Given the number of parameters, which range of parameter values are best supported by the data?

With a classical approach to inference, these estimation problems may be recast as:

1. Test the null hypothesis that the spatial dependence parameters are unnecessary in the model, $H_0 : \quad \theta_\delta = 0$, versus the alternative that they are, $H_A : \quad \theta_\delta \neq 0$, for $\delta = 1, 2, 3, \ldots$.

2. According to some criterion, such as maximising the likelihood, choose the best parameter estimates $\{\hat{\theta}_\delta\}$.

Commonly with spatial or image data, we often have just one sample. The sampling strategy then focusses on 'How wide and diverse an area can we sample?'. The ultimate choice of the approach taken depends on the eventual end-users of the data in fields — such as biogeography, agriculture, medicine and computer science. It is now common, especially with spatial data, for users to prefer a Bayesian approach since this allows them to easily incorporate important and useful prior information; and translates into a very natural data-driven approach.

The basic model for data exhibiting no spatial dependence whatsoever would have just one parameter θ_0, representing overall abundance, or the proportion of sites with a presence. Additional parameters, $\{\theta_\delta\}$, are included to indicate spatial dependence in various directions and at varying degrees of proximity.

In this thesis, I investigate two approaches to model comparison: Bayes factors and extending the hierarchical model. The Bayes factor approach is discussed in more detail below, and this approach is dismissed due to a number of computational issues. An alternative approach to Bayes Factors is to consider the four-tier hierarchical model of equation (5.7). This is the subject of further work within the thesis.

5.7.1 Bayes factors

In the Bayesian paradigm, Bayes factors (BFs) are used to compare pairs of discrete models (Gelman & Meng 1996). Two sets of parameters, which may be of the same or different

dimension, may be compared using the Bayes factor (BF). The BF compares the likelihood of the data under each model represented by the parameter sets.

Let model M_m be the basic model for the data y. We will only be interested in the auto-logistic /Ising model presented in Chapter 4. The parameters $\{\theta_m\}$ are indexed by the model m. Then the Bayes factor for comparing two models, M_0 and M_1 is given by:

$$B_{O1} = \frac{p(y|M_0)}{p(y|M_1)} = \frac{p(M_0|y)}{p(M_1|y)} \Big/ \frac{p(M_0)}{p(M_1)} \tag{5.21}$$

where $p(y|M_m)$ is the likelihood of the data under model M_m (with parameter θ_m); $p(M_m|y)$ is the posterior density of model M_m given the data; and $p(M_m)$ is the user defined prior probability that model M_m is correct. (See Lee (1989) for details.)

Newton & Raftery (1994) suggest a number of methods for estimating the Bayes factor based on using the first definition involving the likeihoods. Each method proceeds by estimating both the numerator and denominator in this definition separately by $p(y|M_m)$ for each M_m. These estimators are based on Monte Carlo simulations from the likelihood of the data and thus can easily be adapted to Markov chain Monte Carlo simulations.

Their estimators are based on two basic estimators. The first, labelled $\widehat{p}_1(\cdot)$, is a harmonic mean of likelihoods for samples from the posterior distributions of $\widetilde{z}_{st}$ and α_k. The second, $\widehat{p}_2(\cdot)$, is an arithmetic mean of likelihoods for samples form the prior distribution of $\widetilde{z}_{st}$ and α_k.

The third and fourth estimators are weighted mixtures of $\widehat{p}_1(\cdot)$ and $\widehat{p}_2(\cdot)$. However, these have theoretical and numerical drawbacks (Gelfand & Dey 1994). A fifth estimator of Gelfand & Dey (1994) improved $\widehat{p}_1(\cdot)$ by importance sampling using an arbitrary probability density function $f(z, \alpha)$ which is a close match to the joint prior $p(z, \alpha)$. Samples are from the posterior distributions $p(\widetilde{z}_{st} \mid \ldots$ and $p(\alpha_k \mid \ldots)$.

In the dingo example a product of IID Bernoulli distributions could be used for an importance sampling function for the prior of $\{\widetilde{z_{st}}\}$ with probabilities of success p_{st} set to posterior estimates $p(\widetilde{z}_{st} = 1 \mid \ldots)$. and a truncated normal for α_k. A truncated Gaussian distribution could be used for α_k or q_k with mean and variance estimated from the posterior distribution $p(q_k \mid \ldots)$.

All five estimators involve summations of exponentiated sums due to the form of $p(z|\theta)$. Thus for this context they suffer from numerical instability and sometimes impractical computational resource requirements.

This approach is detailed in Chapter 7. Suppose that the discretized parameter space for θ contains only 2 models with parameters $\theta_{(A)}$ and $\theta_{(B)}$. The essential feature useful for computing BFs is the posterior odds of selecting M_A over M_B (*i.e.* parameter $\theta_{(A)}$ over $\theta_{(B)}$) that is

$$\begin{aligned} \frac{p(M_A \mid y, z, \alpha)}{p(M_B \mid y, z, \alpha)} &= \frac{p(\theta_{(A)} \mid y, z, \alpha)}{p(\theta_{(B)} \mid y, z, \alpha)} \\ &= \frac{c(\theta_{(B)})}{c(\theta_{(A)})} \exp\left\{(\theta_{(A)} - \theta_{(B)})^{\mathsf{T}} V(z)\right\} \frac{p(\theta_{(A)})}{p(\theta_{(B)})} \end{aligned} \tag{5.22}$$

from equation (5.3).

Then the posterior probability of selecting $\theta_{(A)}$ over $\theta_{(B)}$ is given by

$$p(M_A \mid y, z, \alpha) = \left\{1 + \frac{c(\theta_{(A)})}{c(\theta_{(B)})} \exp\left\{(\theta_{(B)} - \theta_{(A)})^{\mathsf{T}} V(z)\right\} \frac{p(\theta_{(B)})}{p(\theta_{(A)})}\right\}^{-1} \tag{5.23}$$

If more than one model is being considered, say $m = 1, 2, \ldots, M$ then we may use a Metropolis update with a discrete proposal distribution. With a discrete uniform proposal, then the acceptance ratio for switching from model A to B is given by

$$\mathcal{A}(M_B \mid M_A) = \min\left(1, \frac{p(\theta_{(B)} \mid y, z, \alpha)}{p(\theta_{(A)} \mid y, z, \alpha)}\right) \tag{5.24}$$

where the ratio is given above in equation (5.22). Therefore $\mathrm{E}[\,p(\theta_m \mid y, z, \alpha)\,]$ may be estimated by the simulation average of choosing model M_m subject to the usual practical considerations such as convergence, etc. Finally the BF for comparing these 2 models may be computed as

$$\mathcal{B}_{AB} = \frac{\mathrm{E}\left[\,p(\theta_{(B)} \mid y, z, \alpha)\,\right]}{\mathrm{E}\left[\,p(\theta_{(A)} \mid y, z, \alpha)\,\right]} \Big/ \frac{p(\theta_{(B)})}{p(\theta_{(A)})} \tag{5.25}$$

which is the ratio of the posterior to the prior odds of choosing each model.

In addition to the above-mentioned problems with computing Bayes factors for this model, there is the problem of diffuse priors. Due to the lack of expert knowledge in the *dingo* case study (discussed in Chapter 2), it is necessary to apply diffuse priors to capture the range of possible spatio-temporal dingo behaviours. There is documented evidence (Gelman et al. 1995) that Bayes factors are typically difficult to compute when diffuse priors are used, as the ratio of prior distributions then dominates.

5.8 Discussion

Preliminary results for the *dingo* case study provide initialization settings for MCMC simulation of the three-tier model. Of particular use are initial run lengths, starting values for chains, and choices of proposal distributions and their associated parameters. All results in this section condition on the underlying spatio-temporal dependence parameter θ as per equation 5.7. Therefore assessment of posterior densities based on the Bayesian approach are postponed until a full Bayesian approach is implemented in Chapter 7. Posterior distributions are used here to assess design of the MCMC simulation scheme.

This work proves the feasibility of the hierarchical modelling approach under a Bayesian paradigm, but highlights the need for extensions to the model to allow for random θ, and therefore better model choice. The full Bayesian approach is introduced in Section 5.7 during a discussion on model choice. The use of Bayes factors is discussed here but is restricted to pairwise comparison of models. In order to allow a more realistic comparison between a suite of models, a further tier can be added to the modelling hierarchy. In order to implement this a ratio of Normalization constants is required. Methods for estimating these are investigated in Chapter 6, and then the full four-tier hierarchical model is implemented in Chapter 7.

Chapter 6

Inference for the Normalization Constant of Discrete Exponential family distributions

Contents

6.1 Introduction 159
6.1.1 Motivation 159
6.1.2 The problem 160
6.1.3 Review of statistical solutions 162
6.2 Simulation Strategies 163
6.2.1 Independent Importance Sampling Monte Carlo 164
6.2.2 Application to the Autologistic model 164
6.2.3 Improving the Importance Sampling Estimate 165
6.2.4 Variance reduction 166
Antithetic and Control Variates 166
Single *vs* Double simulation 167
6.2.5 Dependent Monte Carlo for ratio of NCs 168
Bridge Sampling 169
6.3 Reverse Logistic Regression 171
6.3.1 Mixture Sample 171
6.3.2 Adjusting variance estimates 173
6.3.3 Computation 174
Overcoming overflow problems 174
Taylor series approximation 175
Additive scaling 175
Multiplicative Scaling 175
Scaling 175
Powering the likelihood 176
6.3.4 Simplifying to two models $M = 2$ 176
6.4 Integrated Mean Canonical Statistic 176
6.4.1 Definition 178

IMCS for Scalar θ 178
Extension to two-dimensional θ 179
Extension to multidimensional θ 180
6.4.2 Numerical Integration of Mean Canonical Statistic 180
Numerical integration: IMCS for scalar θ 180
Path Sampling for Numerical integration: IMCS for two dimensional θ 181
6.4.3 Quadrature Rules 183
6.4.4 Further differentiation 184
6.4.5 Improving integration 185
6.4.6 Error Analysis 186
Comparison 187
6.4.7 Reparameterization 189
6.4.8 Regularization of pairwise IMCS estimates 193
6.4.9 Optimal path 194
6.4.10 Higher order derivatives 195
6.5 Case study 197
6.5.1 Design 197
Extent 197
Prevalence 197
Degree of dependence 198
6.5.2 Base models 199
6.5.3 Simulation Study Results 202
6.5.4 Discussion 205
Relationship between Importance Sampling Monte Carlo (ISMC) and Dependent Monte Carlo (MCMC) 205
Relationship between Integrated Mean Canonical Statistic (IMCS) and Dependent Monte Carlo (MCMC) 213
Reverse Logistic Regression vs Integrated Mean Canonical Statistic 214
6.6 Conclusions 215

6.1 Introduction

To be, or not to be — that is the question:
Whether 'tis nobler in the mind to suffer
The slings and arrows of outrageous fortune
Or to take arms against a sea of troubles,
And by opposing end them?
- *William Shakespeare*, Hamlet, *Act 3, Scene i.*

6.1.1 Motivation

The general problem tackled in this chapter is the evaluation of the normalization constant (NC) or, to be more precise the ratio of two NCs, for a discrete-valued probability distribution defined on a lattice. This evaluation typically involves enumerating the unnormalized form of the distribution over all possible configurations on the lattice. Therein lies the difficulty since straightforward enumeration in problems encountered in practice is often computationally infeasible. Estimation of the ratio of the two NCs can be approached via analytic approximation or simulation. In this work I develop a method which is based on the simulation approach for a specific class of models.

In particular, I consider discrete-valued members of the exponential family of distributions, which offers a wide choice of distributions with support on a lattice. For example, Geyer & Thompson (1992) describe DNA fingerprinting data using an exponential family distribution to model the association of binary data on a lattice. Some other applications were noted in Chapter 2 (Geyer & Thompson 1992, Preisler 1993, Albert & MacShane 1995, Venema 1993, Wolpert & Ickstadt 1995, Denham & Mengersen 1999). In all of these applications, the basic problem is one of inference for a parameter vector θ with a given likelihood $p(x|\theta)$ defined up to normalization. At a more complex level of modelling, discrete distributions on a lattice can be used to represent underlying spatial association. These are exemplified by applications in biogeography where the presence or absence of animals or plants is mapped over a grid (Heikkinen & Högmander 1994, Hoëting, Van Caster & Bowden 1997, Huffer & Wu 1998, Denham & Mengersen 1999, Pettitt & Low Choy 1999, Weir & Pettitt 1999). In addition, in recent years, there has been an increase in the availability and accessibility of Geographic Information Systems (GISs) and satellite imagery. Subsequently this has created an increasing demand for spatio-temporal models for gridded (raster) data.

In almost all of these situations the true binary or discrete process X has a spatial distribution $p(x|\theta)$ which depends on spatial association parameters θ. A hierarchical model can be introduced to model observations Y conditional on the true underlying process X. A particular observation Y_i at site i on lattice L, given the underlying process X_i at the same site i is conditionally independent of all other observations $Y_{L\setminus i}$ given the underlying process X. Thus

$$p(y,x|\theta) = p(y|x)p(x|\theta) = \left\{\prod_i p(y_i|x_i)\right\} p(x|\theta) \tag{6.1}$$

Throughout this chapter I use the notational approach adopted in many texts such as Gelman et al. (1995), where all probability distributions are given the generic notation $p(\cdot)$, with dependence on the random variable being implicitly understood by the context unless specifically stated. All quantities are vectors or matrices unless otherwise stated.

My approach can be considered an alternative to the hidden Gaussian Markov Random Field approach taken by Weir & Pettitt (1999). Their work assumes an underlying con-

tinuous spatial process, modelled by an auto-Gaussian model (Besag 1974). The binary observations are therefore governed by a threshold based on this continuous underlying process; values above the threshold are interpreted as presence; and values below are interpreted as absence. In contrast the approach in this thesis does not assume an underlying continuous spatial process. The problem of estimating the NC of an auto-Gaussian process has already been solved (*e.g.* Baker & Kawashima (1996)).

At the underlying level of the two-layer hierarchical model given in equation (6.1), the distribution $p(x|\theta)$ can be defined either by its full conditional distribution or an unnormalized joint distribution. For binary data on a lattice a suitable form for the conditional distribution is a Markov random field such as the Autologistic model of Besag (1974). The equivalent unnormalized joint distribution is known as a Gibbs or Boltzmann distribution. The equivalence between these conditional and joint expressions has been proven (Dobrushin 1968, Grimmett 1973, Künsch 1983, Glötzl & Rauchenschwandtner 1981). Either way, inference for the spatial process X and its unknown parameter θ requires the evaluation of the normalization constant of $p(x|\theta)$.

The situation described by equation (6.1) commonly arises in image analysis *e.g.* Winkler (1995), Dubes & Jain (1989). In image analysis X can represent the true image whose pattern depends on some parameters θ and Y is the observed 'noisy' image. For pattern recognition problems, the main aim is to estimate the posterior distribution $p(x|y)$, with the unknown parameters θ usually treated on an 'ad-hoc' basis. To this end a Bayesian approach to analysis has been successful and is based on obtaining the posterior distribution of X given Y (Geman & Geman 1984, Besag & Green 1993). Since the focus is on estimating X accurately this is not a direct inferential problem for θ. So in statistical language θ is a 'nuisance' parameter. Analysis often involves an *ad hoc* choice of θ until the estimated image X satisfies both mathematical and non-mathematical measures of 'goodness'. In this chapter I do not set out to focus on extracting the underlying true X although the solution given can be harnessed to solve this problem.

Another application of NC ratios is the computation of Bayes factors, which are used to compare the support for the data observed under different models M_A and M_B with parameters $\theta_A, \theta_B \in \Theta$ respectively. Then the marginal data distribution is given by

$$p(y) = \int_{\Theta} p(y|\theta)p(\theta)d\theta$$

and the Bayes factor for comparing models is defined as (Bernardo & Smith 1994):

$$\mathcal{B}_{AB} = \frac{p(y|M_A)}{p(y|M_B)} = \frac{\int_{\Theta_A} p(y|\theta_A, M_A)p(\theta_A)d\theta_A}{\int_{\Theta_B} p(y|\theta_B, M_B)p(\theta_B)d\theta_B}.$$

This quantity is easily interpreted for a discrete process $p(x|\theta)$.

6.1.2 The problem

This chapter concentrates on the problem of evaluating normalization constants for the exponential family of discrete distributions defined by

$$p(x|\theta) = \frac{h(x,\theta)}{z(\theta)} \tag{6.2}$$

with unnormalized form

$$h(x,\theta) = \exp\{\theta^{\mathsf{T}} V(x)\} \tag{6.3}$$

and normalization constant (NC) given by

$$z(\theta) = \sum_{x \in \Omega} h(x, \theta), \tag{6.4}$$

or more generally

$$z(\theta) = \int h(x, \theta) \mu(dx)$$

where μ is a counting measure or the Lebesgue measure, as required. Here Ω is the sample space of X, is discrete and of high cardinality, and $V(x)$ is a vector sufficient statistic for θ. In the exponential family, the sufficient statistic is also known as the canonical statistic (Cox & Hinkley 1974). The notation $z(\theta)$ is used in a manner similar to $p(\cdot)$, in that it represents the *generic* normalization constant of a distribution obvious in the context. In this case the unnormalized density $h(x, \theta)$ is given by equation (6.2). Throughout, although I emphasise specific exponential family distributions, many of these findings may be applied to other distributions.

As outlined in the previous section, it is the ratio of normalization constants which is often of interest in practical application. NC ratios are easier to work with on the log scale for reasons of scale and numerical stability. Denote the log normalization constant ratio by λ (following the notation of Gelman & Meng (1998)) as follows:

$$\lambda(A, B) = \log \frac{z(\theta_A)}{z(\theta_B)} = \log z(\theta_A) - \log z(\theta_B). \tag{6.5}$$

It is well accepted (Winkler 1995, Besag 1986) that simple enumeration of the normalization constant in these models is far too computationally demanding. For instance, Venema (1993) states that the normalization constant [of binary Markov random fields] is

> *"notoriously intractable ... [and has] only been determined asymptotically for a few regular lattices, amongst others the quadratic and the triangular lattice (Onsager, 1944; Houtabbel, 1950)."*

Its evaluation (asymptotically) has only achieved for a special case of the Ising model with a single parameter, pairwise cliques and isotropic spatial dependence. (See Section 4.3 for more details defining this special case, the simple Ising model.) A routine application of these models to spatial statistics could involve a lattice with at least $L = 1000$ sites. In image analysis an image often comprises at least $L = 1024 \times 1024$ pixels. The enumeration of a binary valued random variable defined on the lattice would require 2^L separate evaluations of the summand.

Historically some of these processes have been investigated via simulation due to the computational problems associated with evaluating the NC. Once Markov Chain Monte Carlo (MCMC) was introduced in Metropolis et al. (1953) as a computational method of obtaining dependent simulations from one of these processes, variations of the algorithm have been devised, each suited to particular conditions. These are discussed in more detail in Section 4.4 with particular reference to the Autologistic distribution. I found the Metropolis-Hastings method to be adequate to the needs of problems such as the *dingo* case study, and discuss it in more detail also in Section 4.4.

Statisticians developed inferential approaches for X and θ in equation (6.2) which all rely on approximations and *avoided* the computation of the normalization constant. These methods included: coding (Besag 1974); pseudo-maximum likelihood (Besag 1975); asymptotic maximum likelihood (Pickard 1976, Pickard 1977*a*, Pickard 1979, Pickard 1987) and

least squares (Possolo 1986*a*). None of these methods requires evaluation of the NC although they all have drawbacks, namely restrictions on lattice size, bias in estimating parameters; limited degree of spatial association or arbitrary choice of cells. See Chapter 4, or Chen (1988) and Possolo (1986*a*), for a review and comparison of their relative merits.

6.1.3 Review of statistical solutions

I now turn to outlining methods from the literature on evaluation of these normalization constants, or ratios of two NCs, using either analytical approximations or simulation methods or a combination of these.

The statistics literature contains at least four main reasons for pursuing research on evaluation of normalization constants and this has resulted in different strategies being developed. The first area of methodological development, for example Geyer (1994, 1996), addresses the general problem of NC ratio estimation and then uses the estimated NC ratio for inference about parameters. The second area, with a recent example being Lewis & Raftery (1997), is also concerned with inference, but uses Bayes factors to choose between 'models' and therefore the associated parameters. A multi-dimensional integration of a likelihood which is equivalent to a NC appears in both the numerator and the denominator of a Bayes factor. Again only NC ratios are required. The third area (Albert & Chib 1995, Carlin & Chib 1995, Green 1995, Madigan & York 1995, George & McCulloch 1993) is focussed on incorporating another level into the modelling hierarchy to allow for 'model' choice. Normalization constants are then required for switching between models, and again only ratios are required. A fourth area closely related to the third commonly arises in a Frequentist context as well as a Bayesian one. Computing conditional likelihoods (Geyer & Thompson 1992), and missing data likelihoods (Gelman & Meng 1998), are both examples of 'Renormalizing' distributions which appear in statistical physics (Binder & Heermann 1988), and depend on estimation of the normalization constant.

In particular, Gelman & Meng (1998) provide a major review of theoretical methods suitable for estimating NC ratios written independently, and in parallel, to this work first published in Low Choy & Pettitt (1997). The method I select for estimating NC ratios with application to the *dingo* case study is called the Integrated Mean Canonical Statistic method. Gelman & Meng (1998) provide a general statistical framework for their Path sampling method, of which IMCS is a special case. I investigate more of the implementation issues for this special case, particularly in the context of a multivariate modelling problem. Their work complements the work in this chapter, examining different elements of the literature as well as different applications. One overlap is the consideration of the work of Ogata & Tanemura (1984) which has proved pivotal in the development of estimators for NC ratios.

Although the application of the normalization constant estimates differ in each of these research areas, the strategies used overlap in various ways. As for many mathematical problems, an analytic approach or a simulation approach to solution can be taken.

A few authors have pursued an analytic approximation strategy by considering Taylor Series expansions for the log NC. These have tended to apply only to continuous state-space distributions, such as spatial point processes. Strauss (1975) used a Taylor Series expansion for the log NC, with coefficients that are the cumulants of the sufficient statistic, then expanding around a value where dependence is absent, thus allowing computation of the first few cumulants. However, this method is only accurate for situations with weak dependence. Gelman et al. (1995) suggest that this method might provide a simple analytic

first approximation.

Tierney & Kadane (1986) gave the Laplace approximation for a Taylor series expansion of the integral of $h(x)$ required to evaluate the NC. Lewis & Raftery (1997) further develop the technique, calling it the 'Laplace-Metropolis Estimator' and use MCMC to estimate either the value of the parameter θ which maximizes h or the multivariate posterior median of θ. The posterior covariance matrix is used to estimate the asymptotic variance matrix defined by the Hessian. They then applied their method to a logistic hierarchical model with random-effects at the underlying error level. These methods do not apply to a discrete state space required for this thesis.

Strategies taking the simulation approach can be grouped as follows. The first strategy type focuses on simulations based on Monte Carlo approaches, and is discussed in Section 6.2. The second strategy, Reverse Logistic Regression, is addressed in Section 6.3. It is highly computationally intensive and was developed by Geyer (1994) in the same general context of 'Normalizing Constant' families. Finally, in Section 6.4 I tailor a method previously only applied to continuous distributions for the discrete exponential family. I call this estimator the Integrated Mean Canonical Statistical (IMCS) estimator to distinguish it from the other estimators considered here. This estimator was independently developed as part of a more general technique called Path Sampling developed in the Statistical Physics literature (Valleau & Card 1972). Low Choy & Pettitt (1997) developed a version of the Path sampling method within the context of an autologistic model. Gelman & Meng (1998) further explore this method and place it within a more general statistical framework. The link between path sampling and IMCS was not recognized previously.

To illustrate discussions I have selected the 3-parameter anisotropic Autologistic model which is introduced in Section 4.3. This model is a useful extension to the isotropic one- or two-parameter versions popular in the literature (Besag 1986, Heikkinen & Högmander 1994, Huffer & Wu 1998). In many cases, approaches to all these strategies have in some way capitalized on MCMC computational methods. MCMC strategies suitable for the models examined in this chapter are discussed briefly in Section 4.4.

Finally Section 6.5 explores the behaviour of each of the estimators in a case study based on an applied statistics problem on dingoes (Pettitt & Low Choy 1999).

6.2 Simulation Strategies

The problem of estimating the NC is essentially one of estimating a sum or integral S of the form:

$$S = \sum_{x \in \Omega} h(x|\theta) \quad \text{or} \quad S = \int h(x|\theta)\mu(dx)$$

where μ is the appropriate measure for X. Ω is the sample space. The NC $z(\theta)$ from the discrete exponential family, including the AL(θ) model, is a sum like S with $h(x|\theta) = \exp\{\theta^T V(x)\}$. In this section we investigate several strategies suggested in the literature for estimating such a sum. They are based on two basic types of Monte Carlo simulations:

- simulations of *independent* identically distributed random variables, using importance sampling; and
- *dependent* simulations from $h(x \mid \theta)$ using Markov Chain Monte Carlo.

The simplest form of Monte Carlo is a special case of Importance Sampling Monte Carlo method introduced in Section 6.2.1 and applied to the autologistic in Section 6.2.2.

Choice of the importance sampling function is further explored in Section 6.2.3. Methods of reducing the variance for this form of Monte Carlo simulation are explored in Section 6.2.4.

Methods based on dependent MC simulations are presented in the next Section 6.2.5. Topics covered include Bartlett's simple expression for the moment generating function, the Geyer-Thompson NC estimator, and the bridge sampling estimator.

6.2.1 Independent Importance Sampling Monte Carlo

A form of the Monte Carlo method uses independent simulations from a distribution $g(\cdot)$ which is 'close' to the distribution of interest $p(\cdot)$. For $g(\cdot)$ to be a good choice it is also required that independent samples may easily be generated from this distribution. Analytic manipulation of the sum (ignoring the dependence on θ) gives:

$$S = \sum_{x\in\Omega} h(x) = \sum_{x\in\Omega} \frac{h(x)}{g(x)} g(x) \equiv \mathrm{E}_g \left[\frac{h(x)}{g(x)} \right] \tag{6.6}$$

where $\mathrm{E}_g[\,\cdot\,]$ is the expectation taken with respect to distribution g. Independent simulations $x^{(1)}, x^{(2)}, \ldots, x^{(T)}$ are generated from the importance sampling distribution g. The sum S can be estimated by using a simulation average

$$\widehat{S} = \frac{1}{T} \sum_{t=1}^{T} \frac{h(x^{(t)})}{g(x^{(t)})}. \tag{6.7}$$

Standard random sample statistical techniques can be used to assess the uncertainty of the estimate $\widehat{S}$.

6.2.2 Application to the Autologistic model

The Simple or Crude Monte Carlo estimator (Hammersley & Handscomb 1964) takes g as uniform. In the case of the binary-valued AL(θ) model this translates to a Bernoulli distribution with probability $p = \frac{1}{2}$ of presence at each site of the lattice. For this simple binary case, we consider a rectangular lattice with L_c columns and L_r rows totalling $L = L_c L_r$ sites. If each simulation $x^{(t)} = (x_1^{(t)}, x_2^{(t)}, \ldots, x_L^{(t)})$ is generated so that the $x_i^{(t)}$ are independent Bernoulli($\frac{1}{2}$) then their distribution is

$$g(x^{(t)}) = \left(\tfrac{1}{2}\right)^L, \qquad x_i^{(t)} \in \{0,1\}, \quad i = 1,\ldots,L, \quad t = 1,\ldots,T$$

so using the definition in equation (6.7) we obtain

$$\widehat{z}(\theta) = \frac{2^L}{T} \sum_{t=1}^{T} h(x^{(t)}|\theta)$$

with $h(x|\theta) = \exp\{\theta^{\mathsf{T}} V(x)\}$ as defined in equation (6.3) and more specifically for the autologistic distribution in equation (4.62). A ratio of NCs can be estimated by the ratio of NC estimators:

$$\widehat{\lambda}_{\mathrm{IS}}(AB) := \log\left\{\frac{\widehat{z}(\theta_A)}{\widehat{z}(\theta_B)}\right\} = \log\left\{\frac{\sum_{t=1}^{T} h(x_A^{(t)}|\theta_A)}{\sum_{t=1}^{T} h(x_B^{(t)}|\theta_B)}\right\}, \tag{6.8}$$

where all elements of $\{x_A^{(t)}\}$ are *iid* Benoulli (p) independently of elements of $\{x_B^{(t)}\}$. Because of the exponential summand h in both the numerator and denominator, this estimate tends

to suffer from poor numerical stability and large variability. We shall demonstrate this below.

Consider $h(x^{(t)}|\theta) = \exp\{\theta^{\mathsf{T}} V(x^{(t)})\}$. For a reasonably large lattice, $V(x^{(t)})$ will have an approximate multivariate normal distribution and therefore $\theta^{\mathsf{T}} V(x^{(t)})$ will have an approximate normal distribution with mean $\mu(\theta)$ and variance $\sigma^2(\theta)$. So $h(x^{(t)}|\theta)$ has approximate mean and variance given by

$$\begin{aligned} \mu_h(\theta) &= \exp\{\mu(\theta) + \tfrac{1}{2}\sigma^2(\theta)\} \\ \sigma_h^2(\theta) &= \exp\{2\mu(\theta) + \sigma^2(\theta)\}(\exp\{\sigma^2(\theta)\} - 1). \end{aligned}$$

Let L_k be the sizes of the neighbourhood sets $\mathcal{N}_k$ defined in equation (4.60). Recall that L is the size of lattice $\mathcal{L}$. Then the mean and variance of $\theta^{\mathsf{T}} V(x^{(t)})$ are given, respectively, by

$$\begin{aligned} \mu(\theta) &= \tfrac{1}{2}\theta_0 L + \tfrac{1}{4}\sum_{k=1}^{K} \theta_k L_k \\ \sigma^2(\theta) &= \tfrac{1}{4}\theta_0^2 L + \tfrac{3}{4}\sum_{k=1}^{K} \theta_k^2 L_k + r(x) \end{aligned}$$

where $r(x)$ comprises terms from the covariance of the components of $V(x)$ (typically a quadratic form in θ with coefficients the order of the L_k.) These aspects are reflected by results given later.

Note that both the bias μ_h and accuracy σ_h^2 depend on $\sigma^2(\theta)$ which in turn is greatly impacted by the size of the neighbourhood and strength of spatio-temporal dependence. The drawback with completely random sampling in this case is that the sum is completely dominated by larger values of h.

We therefore consider methods to improve variance estimation: importance sampling (Section 6.2.3) and variance reduction ideas (Section 6.2.4).

6.2.3 Improving the Importance Sampling Estimate

The importance sampling Monte Carlo estimate is more accurate for functions g which closely approximate p (Hammersley & Handscomb 1964). The difficulty lies in finding a function g to approximate h, from which it is 'simple' to generate independent samples, since it is not the complexity of h which precludes enumeration but high dimensionality of the lattice. We now endeavour to find an appropriate g.

The probability of observing a presence conditional on neighbouring presences on the lattice is:

$$\Pr\{X_i = 1|X_{\mathcal{N}(i)}\} = \frac{\exp\{\theta_0 + \sum_k \theta_k \sum_{j\in\mathcal{N}_k} x_j\}}{1 + \exp\{\theta_0 + \sum_k \theta_k \sum_{j\in\mathcal{N}_k} x_j\}} \tag{6.9}$$

If autodependence is small in a AL(θ) model, then $\theta_k \approx 0$, $k = 1, 2, \ldots, K$. Thus $\pi = \Pr\{X_i = 1|X_{\mathcal{N}(i)}\}$ becomes independent of neighbouring sites i and therefore constant over the entire lattice. This probability π can be expressed in terms of θ_0 as follows:

$$\pi = \frac{\exp\theta_0}{1 + \exp\theta_0} \quad \text{or equivalently} \quad \theta_0 = \log\frac{\pi}{1-\pi} \tag{6.10}$$

As dependence increases, this approximation will become less and less accurate since cross-product terms $\sum_k \theta_k x_i \sum_{j\in\mathcal{N}_k} x_j$ will have more impact on the exponent $\theta^{\mathsf{T}} V(x)$ of h.

With low autodependence, the process X can be well approximated by an independent Bernoulli process $g(x_i) \sim$ i.i.d. Bernoulli(π), where π is the overall lattice marginal probability of a presence. The joint distribution over all sites is

$$g(x^{(t)}|\pi) = \pi^{\sum x_i^{(t)}}(1-\pi)^{L-\sum x_i^{(t)}}$$

Furthermore by using the Autologistic notation $V_0(x) = \sum x_i$ and using equation (6.10) we obtain an importance sampling estimate for the NC $z(\theta)$ by applying equation (6.7)

$$\begin{aligned} \widehat{z}(\theta) &= \frac{1}{T}\sum_{t=1}^{T} \frac{\exp\{\theta^T V(x^{(t)})\}}{\pi^{V_0(x^{(t)})}(1-\pi)^{L-V_0(x^{(t)})}} \\ &= \tfrac{1}{T}\sum_{t=1}^{T} \exp\{\sum_{k=1}^{K}[\theta_k V_k(x^{(t)})] - L\log(1-\pi)\} \end{aligned} \tag{6.11}$$

The simplification follows from noting that the denominator can be rewritten as

$$\left(\frac{\pi}{1-\pi}\right)^{V_0(x)}(1-\pi)^L = \exp\{V_0(x)\theta_0 + L\log(1-\pi)\}.$$

This estimator has essentially the same form, and suffers from similar problems, as the Importance Sampling Estimate based on Bernoulli($\frac{1}{2}$) given in Gelman & Meng (1998). Importance sampling was also investigated by Chen & Shao (1997) and Gelman & Meng (1998) based on *iid* samples. The ratio importance sampling method of Chen & Shao (1997) is essentially equations (6.6)–(6.7).

6.2.4 Variance reduction

Antithetic and Control Variates

The variance of an MC estimator may be reduced using various methods, including the use of antithetic variates or control variates (Hammersley & Handscomb 1964). If moments of a related function are known, then control variates may be constructed to facilitate estimation of another density. For instance, suppose we are interested in estimating $S = \int h(x)\mu(dx)$. The straightforward estimator is $\widehat{h} = \frac{1}{T}\sum_t h(x^{(t)})$ where $\{x^{(t)}\}$ are generated from a uniform distribution. Suppose that the integral $\overline{g} = \int g(x)\mu(dx)$ is tractable, and its estimator is $\widehat{g} = \frac{1}{T}\sum g(x^{(t)})$. Then

$$\mathrm{E}\left[\tfrac{1}{T}\sum(g(x^{(t)}) - \overline{g})\right] = 0$$

so that

$$\tfrac{1}{T}\sum\left[h(x^{(t)}) - (g(x^{(t)}) - \overline{g})\right]. \tag{6.12}$$

also has mean S. The function g is chosen so that $h(x^{(t)}) - (g(x^{(t)}) - \overline{g})$ has smaller variance than $h(x^{(t)})$. Since no analytic expressions for the marginal distributions are known, it is not easy to see a way of choosing a suitable control variate g to reduce the variance of the importance sampling MC estimator.

Single *vs* Double simulation

Another method of reducing the variance would be to consider using the same simulation in both the numerator and denominator of equation (6.13). We investigate the change in variance achieved by doubling the use of simulation variates for general p.

The importance sampling estimate of an NC ratio devised specifically for the autologistic distribution in equation (6.11) can be generalised to any p satisfying equation (6.2). Suppose simulated values $\{x_A^{(t)}\}$ and $\{x_B^{(t)}\}$ are generated independently from importance sampling distributions g_A and g_B, which were selected for their 'closeness' to distributions $p(x|\theta_A)$ and $p(x|\theta_B)$ respectively. Then the NC ratio may be estimated by the expression

$$\widehat{\lambda}_{\text{IS2}}(AB) = \log\left\{\frac{\frac{1}{T}\sum_{t=1}^{T}\psi_A(x_A^{(t)})}{\frac{1}{T}\sum_{t=1}^{T}\psi_B(x_B^{(t)})}\right\} = \log\left\{\frac{\overline{\psi}_A(x_A)}{\overline{\psi}_B(x_B)}\right\}. \tag{6.13}$$

where $\psi_m(x) = \frac{h(x|\theta_m)}{g_m(x)}$ and $\overline{\psi}_m$ is the simulation average $\frac{1}{T}\sum_{t=1}^{T}\psi_m(x_m^{(t)})$.

An option is to use the same set of simulated values in the numerator and denominator of equation (6.13). Thus $x_A^{(t)} = x_B^{(t)} = x^{(t)}$, $\forall t = 1, \ldots, T$ and

$$\widehat{\lambda}_{\text{IS1}}(A, B) = \log\overline{\psi}_A(x) - \log\overline{\psi}_B(x) \tag{6.14}$$

The variance of $\widehat{\lambda}_{\text{IS1}}(A, B)$ therefore involves the variances of each $\log\overline{\psi}_m(x)$ as well as the covariance $\text{Cov}\left[\log\overline{\psi}_A(x), \log\overline{\psi}_B(x)\right]$. Now

$$\begin{aligned}\log\overline{\psi}_m(x) &= \log\left(1 + \frac{\overline{\psi}_m - \mu_m}{\mu_m}\right) + \log\mu_m \\ &\approx \frac{\overline{\psi}_m - \mu_m}{\mu_m} + \log\mu_m\end{aligned}$$

for $\overline{\psi}_m$ close to μ_m and $\mu_m > 0$. Hence the covariance term contains the expression

$$\begin{aligned}\text{Cov}\left[\overline{\psi}_A(x), \overline{\psi}_B(x)\right] &= \frac{1}{n^2}\text{Cov}\left[\sum_{t=1}^{T}\psi_A(x^{(t)}), \sum\psi_B(x^{(t)})\right] \\ &= \frac{1}{n^2}\sum_{t=1}^{T}\text{Cov}\left[\psi_A(x^{(t)}), \psi_B(x^{(t)})\right] \\ &= \frac{\sigma_{\psi_A,\psi_B}}{n}.\end{aligned}$$

So now

$$\begin{aligned}\text{Cov}\left[\log\overline{\psi}_A(x), \log\overline{\psi}_B(x)\right] &\approx \text{Cov}\left[\overline{\psi}_A(x), \overline{\psi}_B(x)\right]\frac{1}{\mu_A}\frac{1}{\mu_B} \\ &= \sigma_{\psi_A,\psi_B}\frac{1}{n}\frac{1}{\mu_A}\frac{1}{\mu_B}.\end{aligned}$$

and

$$\text{Var}\left[\widehat{\lambda}_{\text{IS1}}(A, B)\right] = \text{Var}\left[\log\overline{\psi}_A(x)\right] + \text{Var}\left[\log\overline{\psi}_B(x)\right] - \sigma_{\psi_A,\psi_B}\frac{1}{n}\frac{1}{\mu_A}\frac{1}{\mu_B}.$$

In comparison, the variance of $\widehat{\lambda}_{\text{IS2}}(A, B)$ has no covariance term and is simply

$$\text{Var}\left[\widehat{\lambda}_{\text{IS2}}(A, B)\right] = \text{Var}\left[\log\overline{\psi}_A(x)\right] + \text{Var}\left[\log\overline{\psi}_B(x)\right].$$

Therefore we may conclude that if $\sigma_{\psi_A,\psi_B} > 0$, then the single-simulation estimator has lower variance, and is the better choice. Otherwise the double-simulation estimator is the better choice.

Chen & Shao (1997) develop a variation of the single-simulation option, which in addition has the same importance sampling function in both the numerator and the denominator. They call it the "Acceptance Ratio" method. The selection of an appropriate g requires optimizing the minimum mean square error. Their exposition is restricted to the case where simulation samples $\{x^{(t)}\}$ are independently obtained. The variance results are therefore not appropriate for application to a distribution such as the autologistic where only dependent sampling is currently feasible.

6.2.5 Dependent Monte Carlo for ratio of NCs

An alternative method to the above for estimating the NC ratio is to obtain an analytic expression for the ratio of NCs and then estimate this via dependent MCMC simulations using importance sampling. An early useful result presented by Bartlett (1971) re-expresses the NCs in ratio form allowing Monte Carlo estimation to proceed from a different perspective. The ratio of NCs can be expressed as a moment generating function M of the canonical statistic $V(x)$ based on equations (6.2)–(6.4):

$$\begin{aligned} M_\theta(\phi) &= \mathrm{E}_\theta\left[\exp\{\phi^\mathsf{T} V(x)\}\right] \\ &= \frac{\sum_{x\in\Omega}\exp\{\phi^\mathsf{T} V(x)\}\exp\{\theta^\mathsf{T} V(x)\}}{z(\theta)} \\ &= \frac{z(\phi+\theta)}{z(\theta)} \end{aligned}$$

This is a variation of the result given in Geyer & Thompson (1992), with θ_B written for θ and $\theta_A = \theta_B + \phi$:

$$\begin{aligned} z(\theta_A) &= \sum_{x\in\Omega}\exp\{{\theta_A}^\mathsf{T} V(x)\} \\ &= \sum_{x\in\Omega}\exp\{(\theta_A-\theta_B)^\mathsf{T} V(x)\}.\exp\{{\theta_B}^\mathsf{T} V(x)\} \\ &= z(\theta_B)\sum_{x\in\Omega}\exp\{(\theta_A-\theta_B)^\mathsf{T} V(x)\}p(x|\theta_B) \end{aligned}$$

which yields

$$\frac{z(\theta_A)}{z(\theta_B)} = \sum_{x\in\Omega}\exp\{(\theta_A-\theta_B)^\mathsf{T} V(x)\}p(x|\theta_B). \tag{6.15}$$

Alternatively and more simply this result may be derived from equation (6.7) in general. Suppose the importance sampling function in equation (6.6) is

$$g(x) = p(x|\theta_B) = \frac{h(x|\theta_B)}{z(\theta_B)}.$$

Then

$$z(\theta_A) = \sum_{x\in\Omega}\frac{h(x|\theta_A)}{h(x|\theta_B)}z(\theta_B)p(x|\theta_B) \tag{6.16}$$

so that

$$\frac{z(\theta_A)}{z(\theta_B)} = \mathrm{E}_B\left[\frac{h(x|\theta_A)}{h(x|\theta_B)}\right]. \tag{6.17}$$

where $\mathrm{E}_B[\cdot]$ denotes expectation with respect to the density $p(x|\theta_B)$. Expressing the normalization constant as an expectation with respect to a probability distribution, allows the use of Monte Carlo methods. We can estimate the ratio $z_{AB} = \frac{z(\theta_A)}{z(\theta_B)}$ by using a Monte Carlo estimate based on simulated random variables

$$x^{(t)} \sim p(x|\theta_B) = \frac{\exp\{\theta_B{}^{\mathsf{T}} V(x)\}}{z(\theta_B)}. \tag{6.18}$$

The estimate is given by

$$\widehat{\lambda}_{\mathrm{DMC}}(AB) = \log\left(\frac{1}{T}\sum_{t=1}^{T}\exp\{(\theta_A - \theta_B)^{\mathsf{T}} V(x^{(t)})\}\right)$$

where $\{x^{(t)}, t = 1, \ldots, T\}$ are the simulated observations on a lattice from the distribution $p(x|\theta_B)$. Contrast this to previous situations where $\{x_A^{(t)}\} \sim p(x|\theta_A)$ independently of $\{x_B^{(t)}\} \sim p(x|\theta_B)$. Several simulation methods are available for p being a binary Markov Random Field, but all produce dependent samples as detailed on page 162.

Applying the ergodic theorem to the dependent samples $\{x^{(t)}\} \sim p(x|\theta_B)$ (Geyer & Thompson 1992) yields

$$\widehat{z}_{AB}^{(T)} = \frac{1}{T}\sum_{t=1}^{T}\exp\{(\theta_A - \theta_B)^{\mathsf{T}} V(x^{(t)})\} \xrightarrow{\text{a.s.}} z_{AB} \text{ as } T \to \infty. \tag{6.19}$$

Here the notation $\xrightarrow{\text{a.s.}}$ indicates convergence almost surely, *i.e.* the convergence occurs in the limit as $T \to \infty$ with probability one or, for $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr\left(|T_n - T| \leq \epsilon\right) \to 1.$$

Bridge Sampling

Further inspection of equation (6.16) reveals useful symmetries. Consider the expression

$$\begin{aligned}\sum_{x\in\Omega} h(x|\theta_A)h(x|\theta_B) &= z(\theta_B)\sum_{x} h(x|\theta_A)\frac{h(x|\theta_B)}{z(\theta_B)} \\ &= z(\theta_B)\mathrm{E}_B\left[h(x|\theta_A)\right]. \end{aligned} \tag{6.20}$$

Equivalently this expression, equation (6.20), is also

$$\sum_{x\in\Omega} h(x|\theta_A)h(x|\theta_B) = z(\theta_A)\mathrm{E}_A\left[h(x|\theta_B)\right]. \tag{6.21}$$

Equating equation (6.20) and equation (6.21) gives

$$\widehat{\lambda}_{\mathrm{MW1}}(A, B) = \log\left(\frac{\mathrm{E}_B\left[h(x|\theta_A)\right]}{\mathrm{E}_A\left[h(x|\theta_B)\right]}\right) \tag{6.22}$$

which is a variant of the ordinary Importance Sampling approach, but with samples from $p(x|\theta_A)$ used to estimate $h(x|\theta_B)$ and vice versa. Meng & Wong (1996) generalized this

approach by introducing a "bridge" function $\alpha(x)$, a combination of functions $h(x|\theta_A)$ and $h(x|\theta_B)$, into equation (6.20):

$$\widehat{\lambda}_{\text{MW}}(A,B) = \log\left(\sum_{x\in\Omega} h(x|\theta_A)h(x|\theta_B)\alpha(x)\right) \tag{6.23}$$

changing equation (6.22) to

$$\widehat{\lambda}_{\text{Brdg}} = \log\left(\frac{\text{E}_B\left[\,h(x|\theta_A)\alpha(x)\,\right]}{\text{E}_A\left[\,h(x|\theta_B)\alpha(x)\,\right]}\right). \tag{6.24}$$

Meng & Wong (1996) call this approach "Bridge Sampling". It is also related to the work of Bennett (1976).

A persistent challenge is the optimal choice of α in practice to ensure the 'best' estimate of $\lambda(A,B)$. One option is to minimize the variance of the NC ratio estimator. Assuming equal simulation sizes each from $h(x|\theta_A)$ and $h(x|\theta_B)$, an optimal $\alpha(x)$ which minimizes the relative Mean Square Error is (Meng & Wong 1996):

$$\alpha(x) \propto \frac{1}{\exp\lambda(B,A)h(x|\theta_A) + h(x|\theta_B)}. \tag{6.25}$$

The trouble with this estimator is that it involves the quantity $\lambda(B,A)$ being estimated! An iterative method over steps $\tau = 1,\ldots,\tau_{\max}$ was proposed by Meng & Wong (1996) which the authors describe as being similar to a ratio estimator from independent sampling theory.

$$\exp\widehat{\lambda}_{\text{MWIt}}(A,B)^{(\tau+1)} = \frac{\sum_{t=1}^{T} \frac{h(x_A^{(t)}|\theta_B)}{\exp\widehat{\lambda}_{\text{MWIt}}(A,B)^{(\tau)}h(x_A^{(t)}|\theta_A)+h(x_A^{(t)}|\theta_B)}}{\sum_{t=1}^{T} \frac{h(x_B^{(t)}|\theta_A)}{\exp\widehat{\lambda}_{\text{MWIt}}(A,B)^{(\tau)}h(x_B^{(t)}|\theta_A)+h(x_B^{(t)}|\theta_B)}} \tag{6.26}$$

Then the numerator and denominator of equation (6.24) can be estimated using the straightforward estimators

$$\sum_{t=1}^{T} h(x^{(t)}|\theta_A)\alpha(x^{(t)}) \quad \text{and} \quad \sum_{t=1}^{T} h(x^{(t)}|\theta_B)\alpha(x^{(t)})$$

A particular choice of α "midway" between $h(x|\theta_A)$ and $h(x|\theta_B)$, with support on $\Omega_A \cap \Omega_B$, is

$$\alpha(x) = \frac{h(x|\theta_C)}{h(x|\theta_A)h(x|\theta_B)}.$$

Due to the additivity of the log NC ratio, we have that

$$\lambda(A,B) = \log\left(\frac{z(\theta_B)}{z(\theta_A)}\right) = \log\left(\frac{z(\theta_C)\big/z(\theta_A)}{z(\theta_C)\big/z(\theta_B)}\right) \tag{6.27}$$

which is just a variation of equation (6.50). So applying equation (6.24) gives

$$\widehat{\lambda}_{\text{MID}}(A,B) = \log\text{E}_A\left[\frac{h(x|\theta_C)}{h(x|\theta_A)}\right] - \log\text{E}_B\left[\frac{h(x|\theta_C)}{h(x|\theta_B)}\right]$$

with estimator

$$\widehat{\lambda}_{\text{MID}}(A,B) = \log \sum_{t=1}^{T} \frac{h(x_A^{(t)}|\theta_C)}{h(x_A^{(t)}|\theta_A)} - \log \sum_{t=1}^{T} \frac{h(x_B^{(t)}|\theta_C)}{h(x_B^{(t)}|\theta_B)}$$

where samples are generated according to $x_A^{(t)} \sim p(x|\theta_A)$, $x_B^{(t)} \sim p(x|\theta_B)$, for $t = 1,\ldots,T$. The optimal bridge distribution is then the weighted harmonic mean, which is equivalent to the optimal bridge function from equation (6.25).

6.3 Reverse Logistic Regression

In Geyer (1994, 1996) Reverse Logistic Regression (RLR) was developed to estimate NCs in models having unnormalized density h intractable to estimation via usual techniques, where in contrast, simulations are relatively easy to obtain. Using this method, the NCs are only estimable up to a common constant of proportionality, hence the method can be used when the *ratios* of NCs are required. Situations where Geyer recommends their use include: computation of Bayes factors; importance sampling; and mixture reweighting. Geyer & Thompson (1992) apply the method to the isotropic 1-parameter and 2-parameter Ising models on a 32×32 lattice. This is of comparable size but of more balanced dimensions compared to the *dingo* case study mentioned in Section 6.1. We extend Geyer's results to the general AL(θ) models.

Geyer's method is highly computational, requiring 2 separate stages of simulation: one to build the mixture distribution; and then one for inference which is based on uncoupled (*i.e.* assuming independence) estimating equations for dependent data.

The aim is to obtain log ratios of NCs $\lambda(A,B)$ for unnormalized densities $h(x,\theta_m)$ having different parameters $\{\theta_m : m = 1,\ldots,M\}$. In Section 6.3.1, we first describe the construction of a mixture sampling distribution, with dependent samples drawn from h. We then show how Geyer recast the estimation problem into a framework similar to logistic regression to obtain RLR estimators of NC ratios. Computational issues are important since two stages of simulation are required, one stage for obtaining the samples from each of the mixture components, followed by the next stage for simulating from the mixture.

6.3.1 Mixture Sample

Suppose we can readily simulate from exponential family densities $\{p(x|\theta_m),\ m = 1,\ldots,M\}$ as defined in equations (6.2)–(6.4). For example, dependent samples from the Autologistic distributions AL(θ) may be obtained with a Metropolis-Hastings MCMC sampler, using the algorithm given in Section 4.4. This method requires combining these samples to form a mixture distribution.

First determine models $m = 1,\ldots,M$ which will be used to construct the mixture distribution. Then simulated configurations $x_m^{(t)}$, $t = 1,\ldots,T_m$, $m = 1,\ldots,M$ are generated from within each distribution $p(x|\theta_m)$ separately. The joint distribution of $\{x_m^{(t)}\}$ and $\{\theta_m\}$ is

$$p(x,\theta_m) = \frac{h(x|\theta_m)}{z(\theta_m)} p(\theta_m)$$

and is called a "mixture" distribution by Geyer. Here only fixed θ_m are being considered, hence the prior distribution $p(\theta_m)$ is discrete over m. The prior chosen by Geyer is weighted

by the number of observations obtained from each component distribution

$$p(\theta_m) = \frac{T_m}{|T|} \quad \text{where} \quad |T| = \sum_{m=1}^{M} T_m. \tag{6.28}$$

The marginal distribution of $\{x\}$ is

$$p(x) = \sum_{m=1}^{M} \frac{h(x|\theta_m)}{z(\theta_m)} p(\theta_m) \tag{6.29}$$

which Geyer calls the "mixture" distribution and denotes by $h_{\text{mix}}(x)$.

The trick used in Geyer (1994) was to rewrite the mixture density in a form similar to logistic regression

$$p(x) = \sum_{m=1}^{M} h(x|\theta_m) e^{\eta_m} \tag{6.30}$$

$$e^{\eta_m} = \frac{T_m}{T} \frac{1}{z(\theta_m)}. \tag{6.31}$$

Hence with $h(\cdot)$ in the exponential family, $p(x) = \sum_{m=1}^{M} \exp\{\theta_m^T V(x) + \eta_m\}$. Evidently the logistic parameters η_m do not uniquely determine individual NCs without an additional arbitrary constraint such as the usual sum-to-one or corner constraints used in regression. The NC ratios however are indeed uniquely determined:

$$\lambda_{\text{RLR}}(A, B) = \log T_A - \log T_B + (\eta_B - \eta_A)$$

The probability of observation X being a member of the (random) mixture component A given that it is in the mixture is

$$p(x|\theta_A, \eta) = \frac{h(x|\theta_A) e^{\eta_A}}{\sum_{m=1}^{M} h(x|\theta_m) e^{\eta_m}} \tag{6.32}$$

For example, in MRF models, this probability is

$$p(x|\theta_A, \eta) = \frac{\exp\left\{-\theta_A^{\intercal} V(x) + \eta_A\right\}}{\sum_{m=1}^{M} \exp\left\{-\theta_m^{\intercal} V(x) + \eta_m\right\}} \tag{6.33}$$

Assuming that all simulated observations were obtained independently, or their labels lost, then the RLR log-likelihood is given by:

$$\ell(\eta|x, \theta) = \sum_{m=1}^{M} \sum_{t=1}^{T_m} \log p_m(x_m^{(t)}|\eta) \tag{6.34}$$

If the labels were known, then observations from a mixture component would be correlated because of the way in which they were simulated. If labels are ignored, however, then Geyer's contention is that the observations are independent. Thus this log-likelihood can be referred to as a log pseudo-likelihood since it is the likelihood of the data ignoring underlying dependence. This situation somewhat echoes one which arises in ANOVA when a completely randomized design is assumed when blocks or hierarchical error structure are

in fact present. Although point estimates may be accurate, variance estimates are affected (Box et al. 1978).

The log-likelihood is maximized when the first derivatives are zero:

$$\frac{\partial}{\partial \eta_A}\ell(\eta) = T_A - \sum_{m=1}^{M}\sum_{t=1}^{T_m} p_A(x_m^{(t)},\eta) \equiv 0 \qquad A \in \{1,\ldots,M\} \tag{6.35}$$

and at the maximum, the matrix of second derivatives is negative definite with the ABth entry being

$$\begin{aligned} -\frac{\partial}{\partial \eta_A^2}^2 \ell(\eta) &= \sum_{m=1}^{M}\sum_{t=1}^{T_m} p_A(x_m^{(t)},\eta)\{1 - p_A(x_m^{(t)},\eta)\} \\ -\frac{\partial}{\partial \eta_A \eta_B}^2 \ell(\eta) &= \sum_{m=1}^{M}\sum_{t=1}^{T_m} p_A(x_m^{(t)},\eta) p_B(x_m^{(t)},\eta) \qquad A \neq B \end{aligned} \tag{6.36}$$

The point estimates for $\widehat{\eta}$ can then be obtained using optimization techniques such as the Nelder-Mead simplex algorithm and BFGS Quasi-Newton method with a mixed quadratic and cubic line search procedure. These are implemented in Matlab (The MathWorks, Inc. 1999) as functions `fmins` and `fminu`.

Ignoring the dependence of the underlying samples, initial estimates of the variability of the estimates $\widehat{\eta}$ can therefore be obtained from this observed Information matrix. Furthermore, the standard error of the log ratio of normalization constants is therefore

$$\text{s.e.}\,[\lambda_{\text{RLR}}(A,B)] = \sqrt{\text{Var}\,[\,\eta_A\,] + \text{Var}\,[\,\eta_B\,] - 2\text{Cov}\,[\,\eta_A,\eta_B\,]}$$

where the variances and covariances are taken directly from the inverse of the observed Information matrix in equation (6.36).

Care needs to be taken during calculation. Computations for the dingo example mentioned earlier reached the limits in Fortran90 (on a Dec Alpha Server 2100 with 4× 275 MHz processors running Digital Unix), for maximum double precision numbers. Therefore Matlab was used for computations on a Silicon Graphics Power Challenge platform, a shared memory parallel supercomputer based on MIPS RISC processors. The operating system is IRIX 6 (Unix). On this system double precision figures are stored in scientific format with the limitation on *both* the mantissa and exponent being governed by the machine's maximum integer. We expect such estimates to be smaller than if the dependence assumption were accounted for.

The RLREs can be obtained in the quickest way for mixtures of only two densities ($M = 2$). This allows us to achieve the goal of estimating the ratios of two NCs. The cost, however, of pairwise NC estimates is the lack of internal consistency between pairs.

6.3.2 Adjusting variance estimates

With binary MRF models it is not yet practicable to obtain independent simulations so we are confined to dependent samples, as provided by MCMC. This is unfortunate since the assumption leading to equation (6.34) based on equation (6.29) is inaccurate when samples are obtained from dependent simulations. With dependent samples the approximation to the mixture distribution in equation (6.29) will have misleading variance properties. Equation (6.34) is therefore a pseudo-likelihood for the parameters given the simulated data.

Even though the samples are dependent, the consistency of the estimator is still assured as equation (6.35) has mean zero. Generally, the estimated variance of the point estimates will be underestimated as there is positive dependence within the simulation chains.

6.3.3 Computation

In order to compute the maximum pseudo likelihood estimates of the parameters η and thus of the normalizing constants, it is possible to use general routines to maximise the unconstrained log pseudo-likelihood, given the first and second derivatives.

Another method is to recast the problem of finding where the first derivatives are zero by minimising the sum of squares of the first derivatives. We can use the method of sequential quadratic programming (NAG routine E04UPF) to minimise this sum of squares where the objective function $f(\eta)$ is the sum of squared first derivative of the log-likelihood $f_m(\eta)$ given in equation (6.35)).

$$f(\eta) = \sum_{m=1}^{M} [f_m(\eta)]^2 \quad \text{where} \quad f_m(\eta) = \frac{\partial \ell(\eta; x, \theta)}{\partial \eta_m} \tag{6.37}$$

A linear constraint on the parameters is required to ensure uniqueness. One possibility is the *sum-to-zero* constraint: $\sum_{m=1}^{M} \eta_m = 0$. Another is the *endpoint* constraint where we specify one η_m value (*e.g.* 0) and vary the others.

The computational technique suggested in Geyer (1994) is Newton-Raphson, provided 'care is taken to guard against overflow in the calculations and against overly large steps'. The drawback of this method is that it is only locally convergent, and thus dependent on the starting value.

An alternative method which is globally convergent, albeit slowly, is also cited in Geyer (1994). It requires successive maximization over each parameter, by setting $\zeta_l = e^{-\eta_l}$ and iteratively solving for:

$$\frac{1}{T_l} \sum_{m=1}^{M} \sum_{t=1}^{T_m} \frac{h_l(x_m^{(t)})}{\sum_{k=1}^{M} h_k(x_k^{(t)})/\zeta_k} \tag{6.38}$$

subject to a constraint such as the convenient

$$\zeta_1 = 1 \tag{6.39}$$

Overcoming overflow problems

When computing expression (6.32), there is danger of overflow. This is because the exponents in the numerator and denominator involve expressions with $\theta^{\mathsf{T}} W(x)$ which can range quite widely, for example from -100 to -1000 in the dingo dataset. On adding the logistic regression variables $\{\eta_m\}$, no value of these can offset this wide range, so we are left with a quantity ranging from perhaps $+1000$ to -1000. When exponentiated this can become a very very small number close to zero, and is in danger of being merely set to zero. Therefore it is important to avoid this problem.

Some options considered are: Taylor series approximation to the exponents, factorizing the main exponent by main factors, scaling down by maximum exponent value, and simplifying to the situation where only $M = 2$ models are being considered.

Taylor series approximation

One step towards a solution is to only keep the dominating term in the sum of exponentiated terms in the denominator. When compared to the numerator, there will be a large amount of cancelling.

The Taylor expansion of the exponential is

$$\begin{aligned} e^a &= 1 + a + \frac{a^2}{2!} + \frac{a^3}{3!} + \ldots \qquad \text{for } -\infty < x < \infty \\ &= \sum_{i=0}^{\infty} \frac{a^i}{i!} \end{aligned} \tag{6.40}$$

So for a in the region of $[-100, -1000]$, say, then $e^a \leq e^{-100} = 3.720 \times 10^{-44}$ and $e^a \geq e^{-1000} \leq e^{-709} = 1.217 \times 10^{-308}$

Additive scaling

Computing the derivatives from equation (6.35), requires a summation over iterations t in MCMC chains m of the probabilities $p_r(x_m^{(t)}; \eta)$. These probabilities can become very small, so it is necessary to take some care to avoid underflow. One option is to remove some additive terms in the exponent. For example, by removing the terms $\exp \eta_r$ and $\exp \theta_r^{\mathsf{T}} W(x_1^{(1)})$ from both the numerator and the denominator of the MRF probability in equation (6.33) we obtain:

$$p_l(x_m^{(t)}, \eta) = \frac{\exp\left\{-\theta_l^{\mathsf{T}}\left(W(x_m^{(t)}) + W(x_1^{(1)})\right)\right\}}{\sum_{k=1}^{M} \exp\left\{\left(-\theta_k^{\mathsf{T}} W(x_m^{(t)}) - \theta_l^{\mathsf{T}} W(x_1^{(1)})\right) + (\eta_k - \eta_l)\right\}} \tag{6.41}$$

Multiplicative Scaling

The exponents in the numerator and denominator depend on $-{\theta_m}^{\mathsf{T}} W(x)$ which involve the summation of indicator funcitons over the entire lattice, and so can typically become very large. In turn, this increases the variability in the ratios of their exponentiated versions. For example, in the auto-logistic model applied to the dingo dataset, these have magnitude between -100 and -1000 in a 135×7 lattice with low overall abundance, $P\{X = 1\} = 0.15$.

Scaling

Another method is to simply scale either or both of the numerator or denominator in equation (6.32) using a constant K such that $\frac{1}{K} < 1$, to ensure that the exponential terms are themselves between 0 and 0.1. So for example, in an MRF model, we can adapt equation (6.35) in this way, and the scaled equations are then to be solved.

Furthermore, in the MRF models, we can scale down these exponents by their maximum. Let $\Theta_m(x) = {\theta_m}^{\mathsf{T}} W(x)$ and let $\Theta_{\max}(x) = \max_{m,x} \Theta_m(x)$ be the maximum exponent over all models m and possible data x. Since each element of $W(x)$ is bounded by 0 and $|L|$, $\Theta_{\max}(x) = \max_m \theta^{\mathsf{T}}(|L|, \ldots, |L|)^{\mathsf{T}} = \max_m \sum_\delta \theta_{m,\delta}$.

Powering the likelihood

One method of assisting convergence in the numerical solution of equations (6.29)–(6.34) is to borrow an idea from Simulated Annealing (Geman & Geman 1984). The basic idea is that a function is maximized at the same value as when the same function, raised to a positive power, is maximized. Applied to the maximum likelihood approach, powering the likelihood up or down assists in changing the concentration of the likelihood, particularly at modes.

6.3.4 Simplifying to two models $M = 2$

Geyer's original intention in his papers (Geyer 1994, Geyer & Thompson 1995) was to calculate the ratios of normalizing constants for *a wide range* of models corresponding to vastly different parameter values. In our situation, however, we have found a problem with choosing a wide range of θ values representing the various patterns of spatial association in the auto-logistic framework. The denominator in expression (6.32) can involve exponents of very different orders of magnitude, making it either impossible to compute, or so different in magnitude from the numerator that the ratio is not computable. Therefore, we would first like to explore the simplest case of just calculating the ratio of two normalizing constants for comparing two auto-logistic models with different parameters.

So, we are considering a mixture of MCMC samples from just two unnormalized densities, h_1 and h_2 with corresponding spatial parameters θ_1 and θ_2. Then equation 6.32 becomes:

$$p_m(x,\eta) = \frac{\exp\left\{-\theta_m^{\mathsf{T}} W(x) + \eta_m\right\}}{\exp\left\{-\theta_1^{\mathsf{T}} W(x) + \eta_1\right\} + \exp\left\{-\theta_2^{\mathsf{T}} W(x) + \eta_2\right\}} \tag{6.42}$$

So the RLR log-likelihood becomes

$$\ell_T(\eta; x, \theta) = \sum_{t=1}^{T} \log p_1(x_1^{(t)}, \eta) + \log p_2(x_2^{(t)}, \eta) \tag{6.43}$$

6.4 Integrated Mean Canonical Statistic

In this section I present the method of the Integrated Mean Canonical Statistic, specifically derived for use with exponential family distributions. I shall abbreviate its name to IMCS. IMCS belongs to the same general framework as the Path Sampling (PS) estimator, as presented in Gelman & Meng (1998), where it was developed for application to probability distributions belonging to the more general 'Normalization Constant Family' defined by Geyer (1996). Gelman & Meng (1998) investigate the intricate theoretical relationships between various methods for estimating log NC ratios: variations of importance sampling Monte Carlo; bridge sampling (§ 6.2.5); and path sampling. The selection of optimal importance sampling functions and the optimal path is emphasised, based on variance minimization arguments. Applications used to illustrate the theory were distributions with a less complex structure than the Autologistic. I find that the optimal choices for importance sampling functions or for path sampling are not necessarily feasible options for distributions such as the Autologistic, so investigate other alternatives. This work aims to find workable solutions to the problem of estimating the normalization constant for exponential families, particularly high dimensional sample space and complex distributions such as the Autologistic.

IMCS is also related to the approach of Strauss (1986) and suggestions of Geyer & Thompson (1992), where it was applied to other *continuous* exponential family distributions. Strauss (1986) simulated from $p(x \mid \theta)$ using Markov chain Monte Carlo. For each simulation, $\mathrm{E}_\theta [\, V(x) \,]$ was estimated, and the resulting surface was smoothed over the regular grid of θ values. Diggle & Gratton (1984) took a similar approach but used independent Monte Carlo.

There is a link to work in spatial point processes, such as Diggle, Fiksel, Grabarnik, Ogata, Stoyan & Tanemura (1994) and Ogata & Tanemura (1984). These papers focus on pairwise interaction point processes, the continuous counterparts of discrete Markov random field models. The methods investigated were analytic approximations, based on virial expansions or Padé approximants to Monte Carlo samples as mentioned in Section 4.3.5, importance sampling Monte Carlo, or the Tajacs-Fiskel method, of which maximum pseudo-likelihood is a special case. The last approach applies in general to Gibbs distributions and is based on a result linking the overall prevalence to a weighted expectation of any function of the random variable X. The applications they consider are all continuous processes, which enables them to take analytic shortcuts which are not available for discrete distributions.

In Section 6.4.1, I define the Integrated Mean Canonical Statistic estimator of the log NC ratio. This is achieved in stages, starting with scalar θ, then proceeding to two-dimensional θ, and finally arriving at multi-dimensional θ. An important issue also considered in this section is the choice of an appropriate path for multi-dimensional θ.

Two steps are required in evaluation of equation (6.47) or equation (6.52). The first step is computation of the mean canonical statistic. Simulation can feasibly provide estimates; these are discussed in Section 6.4.1. The second step in evaluating the log NC ratio is numerical integration of the mean canonical statistic. Methods for achieving this are examined in Sections 6.4.2 for scalar θ. For vector θ, path sampling is introduced in Section 6.4.1, and optimal choices for the path considered in Section 6.4.9.

The next sections discuss various computational issues. I first note that once the mean canonical statistic has been estimated then the problem of estimating the log NC ratio can be viewed from two different perspectives. The first view is based on integration, as described by equation (6.47). Here the focus is on fitting an accurate surface to $\mathrm{E}_\theta [\, V(x) \,]$, *prior to* integration. The mean canonical statistic is not available in closed form for the $\mathrm{AL}(\theta)$ distribution, so numerical approximations are required. These are based on evaluations of the function at various values of θ. A simple approximation to the integrand $\mathrm{E}_\theta [\, V(x) \,]$ is piecewise linear or quadratic, as assumed for simple quadrature rules. These rules are discussed in detail in Section 6.4.3. I find that for my purposes these quadrature rules are quite adequate. For situations where more accuracy is required, more sophisticated integration can be achieved by fitting a more complex surface, as discussed briefly in Section 6.4.5.

The second view is based on differentiation of the log NC as depicted by equation (6.46). Here the essential problem is that I am trying to find the height *differences* of a function, the log normalization constant, given *estimates* of its *partial derivatives*, which are equivalent to the expected value of the canonical statistic. Estimation of the log NC ratio can therefore be achieved by improving the estimation of the derivatives, and is discussed in Section 6.4.4.

When the vector pair θ_A and θ_B differ in more than one component then the log NC ratio may be estimated by reparameterization of θ so that the resulting parameters ϕ differ in only one component. This greatly simplifies computations. Reparameterization is discussed in Section 6.4.7.

Although several pairwise NC ratios can be estimated separately they are all based on

the same underlying NCs. As shown in Section 6.4.8, under certain assumptions these NCs can be estimated (up to a constant of proportionality) via a process of regularization achieved through linear regression.

An error analysis of the quadrature approximations is addressed in Section 6.4.6. A theoretical study of this accuracy is undertaken in Section 6.4.6. Since appropriate analytic results are unavailable for the AL(θ) model, the behaviour of the quadrature rule estimates is investigated for an independent counterpart, the Poisson distribution.

Finally extensions to the IMCS based on higher order derivatives are explored in Section 6.4.10.

6.4.1 Definition

IMCS for Scalar θ

The name "Integrated Mean Canonical Statistic" reflects the operations used by this method to estimate a logged ratio of normalization constants: the **mean** of the **canonical statistic** $V(x)$, from the distribution of interest $p(x \mid \theta)$ defined in equation (6.2), is **integrated**.

The IMCS estimator is based on a simple relationship (Ripley 1988) that is derived from equations (6.2)–(6.4):

$$\begin{aligned}
\frac{\partial}{\partial \theta_k} z(\theta) &= \frac{\partial}{\partial \theta_k} \int_{x \in \Omega} \exp\{\theta^{\mathsf{T}} V(x)\}\, \mu(dx) \\
&= \int_{x \in \Omega} \frac{\partial}{\partial \theta_k} \exp\{\theta^{\mathsf{T}} V(x)\}\, \mu(dx) \\
&= \int_{x \in \Omega} V_k(x) \exp\{\theta^{\mathsf{T}} V(x)\}\, \mu(dx) \\
&= z(\theta) \mathrm{E}_\theta [\, V_k(x) \,]. \qquad (6.44)
\end{aligned}$$

All derivatives of $z(\theta)$ exist and are continuous in the Exponential Family and the range of integration does not depend on θ. Hence the progression to the third step in equation (6.44) above is legal. Substituting the identity

$$\frac{\partial}{\partial \theta_k} \log z(\theta) = \frac{\partial}{\partial \theta_k} z(\theta) / z(\theta) \qquad (6.45)$$

in equation (6.44) gives a series of simultaneous equations based on the first order partial derivatives for each component θ:

$$\begin{aligned}
\frac{\partial}{\partial \theta_0} \log z(\theta) &= \mathrm{E}_\theta [\, V_0(x) \,] \\
\frac{\partial}{\partial \theta_1} \log z(\theta) &= \mathrm{E}_\theta [\, V_1(x) \,] \qquad (6.46) \\
&\vdots \\
\frac{\partial}{\partial \theta_K} \log z(\theta) &= \mathrm{E}_\theta [\, V_K(x) \,].
\end{aligned}$$

For scalar θ rearranging equation (6.46) yields an expression for the difference of two log NCs in terms of a definite integral of $\mathrm{E}_\theta [\, V(x) \,]$:

$$\lambda(A, B) = \log \frac{z(\theta_A)}{z(\theta_B)} = \log z(\theta_A) - \log z(\theta_B) = \int_{\theta_B}^{\theta_A} \mathrm{E}_\theta [\, V(x) \,]\, d\theta. \qquad (6.47)$$

A more general formulation of equations (6.44)–(6.46) was presented in Gelman & Meng (1998) for scalar θ, written in terms of h as follows:

$$\frac{\partial}{\partial\theta}\log z(\theta) = \mathrm{E}_\theta\,[\,V(x)\,] \quad \text{with} \quad V(x) = \frac{\partial}{\partial\theta}\log h(x\,|\,\theta). \tag{6.48}$$

It is a special property of the exponential family that V depends only on x, due to the linear contribution of θ.

Extension to two-dimensional θ

I now consider extending this method to multidimensional θ. For pedagodic reasons, it is beneficial to first consider the simple two-dimensional case, and then expand later to the more general framework of path sampling.

Simultaneously solving the partial first order differential equations in equation (6.46) requires no extra effort, in addition to equation (6.47), for pairs of parameters θ_A and θ_B which differ in precisely one component. Consider a 'discretized' restricted space for the parameter θ, with each component θ_k taking on one of N_k distinct values in $\Theta_k = \{\theta_k^{(1)}, \theta_k^{(2)}, \ldots, \theta_k^{(N_k)}\}$. For instance, with dimension $K = 2$, parameters θ take on values in the full space $\Theta = \Theta_0 \times \Theta_1 \times \Theta_2$. Call two parameter values θ_A and θ_B "adjacent" on such a discrete sample space Θ if they differ in one single component k^*. That is $\theta_{Ak} = \theta_{Bk}$ for $k \neq k^*$ and $\theta_{Ak} \neq \theta_{Bk}$ for some $k = k^*$. Thus the log NC ratio between adjacent parameters θ_A and θ_B can be obtained using equation (6.47) by integrating the mean value of the kth component of the canonical statistic $\mathrm{E}_\theta\,[\,V_k(x)\,]$ over the interval defined by the kth components of the parameters θ_{Ak} and θ_{Bk}:

$$\lambda(A,B) = \int_{\theta_{Bk}}^{\theta_{Ak}} \mathrm{E}_\theta\,[\,V_k(x)\,]\,d\theta. \tag{6.49}$$

For pairs θ_A and θ_B which differ in more component, one may take advantage of the additivity of the log NC ratios. By definition,

$$\log\frac{z(\theta_A)}{z(\theta_B)} = \log\frac{z(\theta_A)}{z(\theta_C)} + \log\frac{z(\theta_C)}{z(\theta_B)}$$

so that

$$\lambda(A,B) = \lambda(A,C) + \lambda(C,B). \tag{6.50}$$

Thus if θ_A, θ_B are two-dimensional, then a single "bridge" parameter θ_C may be identified which differs in only one dimension compared to each of θ_A, θ_B. Parameters $\theta_D, \theta_E, \ldots$ may be introduced as necessary. The log NC ratio $\lambda(A,B)$ may then be estimated using equation (6.50). The bridge parameter θ_C echoes the purpose of the bridge importance sampling function used in Section 6.2.5. There are, of course, two such "bridge" parameters to choose from. On further contemplation, however, it becomes apparent that selecting a series of bridges, via other parameters θ_D, θ_E, is just one way of constructing a path between θ_A and θ_B to facilitate computation of equation (6.47). There is no reason why a path must progress in a stepwise fashion, analagous to the Manhattan city block approach, such as that dictated by the series of one-step bridges. A direct route, or some other route which satisfies optimality criteria, may provide a "better" estimate of λ.

This process of selecting an appropriate path, even an optimal path, was presented in the more general framework of *Path Sampling* by Gelman & Meng (1998).

Extension to multidimensional θ

Extending equation (6.47) to multidimensional θ can therefore be achieved by selecting an appropriate *path* for applying the integration. The following is adapted from Gelman & Meng (1998, p166). Consider the problem of estimating $\lambda(A, B)$. Define a continuous path from θ_A to θ_B using path index $m \in [0, M]$ by

$$\theta(m) = \left[\begin{array}{cccc} \theta_0(m) & \theta_1(m) & \ldots & \theta_K(m) \end{array} \right]^{\mathsf{T}} \quad \text{with} \quad \theta(0) = \theta_A,\ \theta(M) = \theta_B.$$

Use standard notation to denote partial derivatives along the path:

$$\dot{\theta}_k(m) = \frac{\partial}{\partial m}\theta_k(m), \quad k = 1, \ldots, K.$$

Then equation (6.47) may be extended as follows:

$$\begin{aligned} \lambda(A, B) &= \int_{\theta_B}^{\theta_A} \mathrm{E}_{\theta(m)} \left[\frac{\partial}{\partial m} \log h(x \,|\, \theta(m)) \right] dm \\ &= \int_{\theta_B}^{\theta_A} \sum_{k=1}^{K} \dot{\theta}_k(m)\, \mathrm{E}_{\theta(m)} \left[V_k(x) \right] dm \end{aligned} \tag{6.51}$$

although $V_k(x)$ does not depend on $\theta(m)$ in the family of distributions defined by equations (6.2)–(6.4).

When full integration along the path is not feasible, an estimator for λ may be constructed by sampling along the path. One option is to use uniform (random) sampling along the path to provide a Monte Carlo path estimate. Given uniformly distributed samples $\{m_t,\, t = 1, \ldots, T\}$ along the path, and simulations $x^{(t)} \sim p(x \,|\, \theta(m_t))$, the Path sampling estimate of the log NC ratio is:

$$\widehat{\lambda}_P(A, B) = \frac{1}{T} \sum_{t=1}^{T} \left[\sum_{k=1}^{K} \dot{\theta}_k(m_t)\, V_k(x^{(t)}) \right] \tag{6.52}$$

where $\mathrm{E}_{\theta(m)} \left[V_k(x) \right]$ may be approximated by the single realization $x^{(t)}$.

The expression for the NC ratio presented in equation (6.47) requires evaluation of the marginal mean $\mathrm{E}_\theta \left[V(x) \right]$ of the canonical statistic of the distribution $p(x \,|\, \theta)$. Instead of using the single realization approach above, dependent simulations of $V_k(x)$ may be obtained from MCMC simulations, and used to estimate its moments.

6.4.2 Numerical Integration of Mean Canonical Statistic

I will first consider various approaches to obtaining $\lambda(A, B)$ from equation (6.47) for scalar θ and then extend to vector θ. An account is given in Gelman & Meng (1998).

Numerical integration: IMCS for scalar θ

In equation (6.52) integration was carried out using a Monte Carlo approximation. Here we consider using quadrature rules. Using quadrature requires estimation of $\mathrm{E}_\theta \left[V(x) \right]$ at fixed grid of points of θ, say $\theta_A = \theta_1, \ldots, \theta_M = \theta_B$ giving estimates v_m, $m = 1, \ldots, M$. Estimation of $\lambda(A, B)$ can be achieved by a quadrature rule, such as the trapezoidal rule

$$\widehat{\lambda}_T(A, B) = \sum_{m=1}^{M-1} \tfrac{1}{2}(\theta_{m+1} - \theta_m)(v_m + v_{m+1}).$$

Equally spaced θ_m's ensures that $\Delta\theta = \theta_{m+1} - \theta_m, \forall m$, so

$$\widehat{\lambda}_T(A,B) = \frac{\Delta\theta}{2}(v_1 + 2v_2 + \ldots + 2v_{M-1} + v_M). \tag{6.53}$$

The trapezoidal rule has an advantage over other quadrature rules, such as Simpson's, since it is additive for intermediate calculations of the log NC ratio λ:

$$\widehat{\lambda}_T(A,m) + \widehat{\lambda}_T(m,B) = \widehat{\lambda}_T(A,B), \quad \text{for} \quad m = 1,\ldots,M.$$

This property proves to be convenient when extending to the case of vector θ and so-called 'path' sampling.

The Monte Carlo rule arises from the result that

$$\begin{aligned} \lambda(A,B) &= \int_{\theta_A}^{\theta_B} \mathrm{E}_\theta\,[\,V(x)\,]\,d\theta \\ &= \int_{\theta_A}^{\theta_B} \frac{\mathrm{E}_\theta\,[\,V(x)\,]}{g(\theta)} g(\theta) d\theta \end{aligned} \tag{6.54}$$

where $g(\theta)$ is a density with support on $[\theta_A, \theta_B]$. If $\theta_1,\ldots,\theta_m$ are generated according to $g(\theta)$ and v_m is the estimate of $\mathrm{E}_\theta\,[\,V(x)\,]$ at θ_m then $\lambda(A,B)$ can be estimated by

$$\widehat{\lambda}_g(A,B) = \frac{1}{M}\sum_{m=1}^{M} \frac{v_m}{g(\theta_m)}. \tag{6.55}$$

These two approaches are in fact equivalent. By writing $\widehat{\lambda}_T(A,B)$ as

$$\tfrac{1}{2}(\theta_2 - \theta_1)v_1 + \sum_{m=2}^{M-1}\left[\tfrac{1}{2}(\theta_m - \theta_{m-1} + \theta_{m+1} - \theta_m)v_m\right] + \tfrac{1}{2}(\theta_M - \theta_{M-1})v_M \tag{6.56}$$

it becomes apparent that $\widehat{\lambda}_g$ is equal to $\widehat{\lambda}_T$ when the Monte Carlo importance sampling distribution is given by:

$$g(\theta_m) = \begin{cases} \frac{2}{M}(\theta_{m+1} - \theta_{m-1})^{-1} & , \quad 2 \le m \le M-1 \\ \frac{2}{M}(\theta_2 - \theta_1)^{-1} & , \quad m = 1 \\ \frac{2}{M}(\theta_M - \theta_{M-1})^{-1} & , \quad m = M \end{cases} \tag{6.57}$$

This point is briefly noted by Gelman & Meng (1998, §2.3). Hence numerical integration via the trapezoidal rule of equation (6.53) is equivalent to Monte Carlo integration, provided that the sampling distribution is almost uniform when the difference between θ_ms is constant, *i.e.* $\theta_{m+1} - \theta_m = \Delta\theta$.

Path Sampling for Numerical integration: IMCS for two dimensional θ

The results for scalar θ can be extended to vector θ using the idea of 'path sampling' developed by Gelman & Meng (1994) and Gelman & Meng (1998). Let us investigate the two-dimensional parameter $\theta = (\theta_1, \theta_2)$ over a square grid of equally spaced values: $\{(\theta_1^{(a)}, \theta_2^{(b)});\ a = 1,\ldots,M;\ b = 1,\ldots,M\}$. Consider opposite corners of the grid $\theta_A = (\theta_1^{(1)}, \theta_2^{(1)})$ and $\theta_B = (\theta_1^{(M)}, \theta_2^{(M)})$, and another corner $\theta_C = (\theta_1^{(M)}, \theta_2^{(1)})$. A possible 'path'

from θ_A to θ_B in Θ-space can travel via θ_C. The NC ratios between the θ pairs are related by

$$\log \frac{z(\theta_A)}{z(\theta_B)} = \log \frac{z(\theta_A)}{z(\theta_C)} + \log \frac{z(\theta_C)}{z(\theta_B)}$$

so that

$$\lambda(A,B) = \lambda(A,C) + \lambda(C,B).$$

The NC ratio $\lambda(A,C)$ involves constant $\theta_2 = \theta_2^{(1)}$ and θ_1 changing from $\theta_1^{(1)}$ to $\theta_1^{(M)}$, with corresponding canonical statistic $V_1(x)$. The estimate $\widehat{\lambda}_T(A,C)$ is therefore a one-dimensional integral over θ_1. Similarly $\lambda(C,B)$ involves changes only in θ_2, and can be estimated by $\widehat{\lambda}_T(C,B)$ with $V_2(x)$ being the canonical statistic involved.

Alternatively one could re-parameterize so that

$$\begin{aligned} \phi_1 &= \tfrac{1}{\sqrt{2}}(\theta_2 - \theta_1) \\ \phi_2 &= \tfrac{1}{\sqrt{2}}(\theta_2 + \theta_1). \end{aligned} \tag{6.58}$$

This maintains an orthonormal transformation. In ϕ coordinates θ_A and θ_B become

$$\begin{aligned} \phi_A &= \left(\tfrac{1}{\sqrt{2}}(\theta_2^{(1)} - \theta_1^{(1)}), \tfrac{1}{\sqrt{2}}(\theta_2^{(1)} + \theta_1^{(1)})\right) \\ \phi_B &= \left(\tfrac{1}{\sqrt{2}}(\theta_2^{(M)} - \theta_1^{(M)}), \tfrac{1}{\sqrt{2}}(\theta_2^{(M)} + \theta_1^{(M)})\right). \end{aligned} \tag{6.59}$$

For equal $\theta_1^{(m)} = \theta_2^{(m)}$ these become

$$\begin{aligned} \phi_A &= \left(0, \tfrac{1}{\sqrt{2}}(\theta_2^{(1)} + \theta_1^{(1)})\right) \\ \phi_B &= \left(0, \tfrac{1}{\sqrt{2}}(\theta_2^{(M)} + \theta_1^{(M)})\right). \end{aligned} \tag{6.60}$$

In ϕ coordinates ϕ_A and ϕ_B only differ in their values of ϕ_2. To apply the trapezoidal rule $\widehat{\lambda}_T(A,B)$ to the line joining ϕ_A to ϕ_B, I note that the canonical statistic is $\frac{1}{\sqrt{2}}(V_1(x) + V_2(x))$.

Suppose estimated values of $V_1(x)$ and $V_2(x)$ at $(\theta_1^{(a)}, \theta_2^{(b)})$, $a,b = 1,\ldots,M$ are $v_1^{(ab)}, v_2^{(ab)}$, $a,b = 1,\ldots,M$ then the trapezoidal rules gives an estimate for this minimum euclidean distance path:

$$\begin{aligned} &\widehat{\lambda}_{T,Euc}(A,B) \\ &= \textstyle\sum_{m=1}^{M-1} \frac{1}{2}\left(\phi_2^{(m+1)} - \phi_2^{(m)}\right)\frac{1}{\sqrt{2}}\left(v_1^{(mm)} + v_2^{(mm)} + v_1^{(m+1,m+1)} + v_2^{(m+1,m+1)}\right) \\ &\quad \cdot\left(v_1^{(m+1,m+1)} + v_2^{(m+1,m+1)}\right) \\ &= \textstyle\sum_{m=1}^{M-1} \frac{1}{2}\frac{1}{\sqrt{2}}\left(\theta_1^{(m+1)} - \theta_1^{(m)} + \theta_2^{(m+1)} - \theta_2^{(m)}\right) \\ &\quad \cdot\frac{1}{\sqrt{2}}\left(v_1^{(mm)} + v_2^{(mm)} + v_1^{(m+1,m+1)} + v_2^{(m+1,m+1)}\right) \end{aligned}$$

and for equally spaced θ's

$$\begin{aligned} &= \textstyle\sum_{m=1}^{M-1} \frac{1}{2}\frac{1}{2}2\Delta\theta\left(v_1^{(mm)} + v_2^{(mm)} + v_1^{(m+1,m+1)} + v_2^{(m+1,m+1)}\right) \\ &= \tfrac{\Delta\theta}{2} \sum_{m=1}^{M-1}\left(v_1^{(mm)} + v_1^{(m+1,m+1)}\right) + \frac{\Delta\theta}{2}\sum_{m=1}^{M-1}\left(v_2^{(mm)} + v_2^{(m+1,m+1)}\right) \\ &= \tfrac{\Delta\theta}{2} \left(v_1^{(11)}) + 2v_1^{(22)} + \ldots + 2v_1^{(M-1,M-1)} + v_1^{(MM)}v_2^{(11)}\right) \\ &\quad +2v_2^{(22)} + \ldots + 2v_2^{(M-1,M-1)} + v_2^{(MM)}\Big) \end{aligned}$$

Compare this estimate with the trapezoidal rule applied to $[\theta_A, \theta_C]$ and $[\theta_C, \theta_B]$, which is essentially a Manhattan City Block measure.

$$\begin{aligned}\widehat{\lambda}_{T,Man}(A,B) &= \widehat{\lambda}_T(A,C) + \widehat{\lambda}_T(C,B) \\ &= \frac{\Delta\theta}{2}\sum_{m=1}^{M-1}\left(v_1^{(m1)} + v_1^{(m+1,1)}\right) + \frac{\Delta\theta}{2}\sum_{m=1}^{M-1}\left(v_2^{(m1)} + v_2^{(m+1,1)}\right) \\ &= \frac{\Delta\theta}{2}\left(v_1^{(11)} + 2v_1^{(21)} + \ldots + 2v_1^{(M-1,1)} + v_1^{(M1)}\right. \\ &\quad \left. + v_2^{(M1)} + 2v_2^{(M2)} + \ldots + 2v_2^{(M,M-1))} + v_2^{(MM)}\right)\end{aligned}$$

These ideas can obviously be extended to more than two dimensions.

The two estimates involve the same number of terms and coefficients. The difference is that evaluating $\widehat{\lambda}_{T,Euc}(A, B)$ requires estimation of $\mathrm{E}[V_1(x)]$ and $\mathrm{E}[V_2(x)]$ for each value of θ whereas evaluation of $\widehat{\lambda}_{T,Man}(A, B)$ requires estimation of either ($\mathrm{E}[V_1(x)]$ or $\mathrm{E}[V_2(x)]$) for each value of θ except for $(v_1^{(M1)}, v_2^{(M1)})$. Questions that now arise are:

1. For judging the different paths, what criteria are appropriate? Numerical or statistical accuracy?
2. Are there optimal paths based on either numerical or statistical criteria?
3. What are practical approaches to estimating the log ratio NC which have reasonable statistical and numerical properties?
4. Of various quadrature rules, which are better than others for implementation of quadrature estimation of $\lambda(A, B)$?
5. Is averaging over different paths a viable alternative to finding the optimal path, which is impractical? (In physics this is known as Feynman's path integral.)

In answer to the fourth question, various quadrature rules are compared in Sections 6.4.3, 6.4.6 and 6.4.6. These sections also address the first question on combining numerical and statistical measures of accuracy when quadrature techniques are used, in particular an error analysis is given in Section 6.4.6. The third question on practical methods to estimating the log NC are examined in a number of sections. The sections on quadrature mentioned above are relevant, as are Sections 6.4.5 and 6.4.4 which address different approaches to integration or the dual problem of differentiation. Section 6.4.10 considers one extension of this approach obtained by considering higher order derivatives. Section 6.4.7 further addresses the question of practicality by reducing the complexity of integration via reparameterization.

The fifth question on whether averaging over paths is useful is considered under the heading of Regularization in Section 6.4.8. This leads on to the second question on optimal paths, which is discussed in Section 6.4.9.

6.4.3 Quadrature Rules

A simple numerical approach to integration is to apply quadrature rules. Three elementary choices are the trapezoidal rule (T), the rectangular rule (R) and their combined form Simpson's rule (S). In order for these rules to be accurate, the integrand needs to be at worst quadratic. It is not unreasonable to assume here that the canonical statistic surface is

approximately linear over a sufficiently small area between the parameter values of interest. For simplicity write $\overline{V}_{Ak}$ for $\mathrm{E}_{\theta_A}[V_k(x)]$. To numerically find $\int_{\theta_B}^{\theta_A} \mathrm{E}_\theta[V_k(x)]\,d\theta$ one can use an M point quadrature rule Q_M where Q is one of S, T, R. For example, the 2-point trapezoidal rule is:

$$\mathsf{T}_2(\log z_{AB}) = \frac{1}{2}\left(\overline{V}_{Ak} + \overline{V}_{Bk}\right)(\theta_{Bk} - \theta_{Ak})$$

This simple rule requires two point estimates of the canonical statistic evaluated at the parameter values of interest θ_A and θ_B.

The 1-point rectangular rule is:

$$\mathsf{R}_1(\log z_{AB}) = \frac{1}{2}\left(\overline{V}_{Ck}\right)(\theta_{Bk} - \theta_{Ak})$$

where $\overline{V}_{Ck}$ is the kth component of the mean canonical statistic from the model with parameter value midway between that of A and B, $\theta_C = \frac{1}{2}(\theta_A + \theta_B)$. Numerically, the rectangular rule is more precise than the Trapezoidal rule when the first five derivatives exist (Forsythe, Malcolm & Moler 1977). The practical difficulty is that it requires estimation of $\mathrm{E}_\theta[V(x)]$ *between* and not *at* the θ values of interest, and so is not the most efficient use of simulations corresponding to each parameter value. This rule is obviously not well suited to applications where $\log z_{AB}$ is required for given θ_A and θ_B.

The integral obtained from applying Simpson's rule requires more point estimates of the canonical statistic over the desired interval, but gives higher precision. The numerical precision quadruples every time the number of subintervals used is doubled (Forsythe et al. 1977). It is a combination of the best features of the trapezoidal and rectangular rules, designed to reduce the numerical error in the approximation. The 3-point version of Simpson's rule combines the 1-point Rectangular rule and the 2-point Trapezoidal rule:

$$\mathsf{S}_3(\log z_{AB}) = \tfrac{1}{6}\left(\overline{V}_{Ak} + 2\overline{V}_{Ck} + \overline{V}_{Bk}\right)(\theta_{Bk} - \theta_{Ak})$$

Again a drawback of Simpson's rule is that evaluation of $\log z(\theta)$ is required at points intermediate to those of interest.

6.4.4 Further differentiation

For $z(\theta)$ all derivatives exist and are continuous, so that an infinite power series expansion for $\log z(\theta)$ is appropriate. Equations (6.46), (6.72) and (6.73) give expressions so that derivatives can be estimated from simulations. Exploring this idea one can write $\log z(\theta)$ as a quadratic in θ

$$\log z(\theta) = a + \sum_{k=0}^{K} b_k\theta_k + \sum_k \sum_l c_{kl}\theta_k\theta_l + r(\theta), \quad (c_{kl} = c_{lk}),$$

where $r(\theta)$ is the remainder term absorbing higher order terms in θ. Differentiating the power series gives

$$\begin{aligned}
\frac{\partial}{\partial\theta_k}\log z(\theta) &= b_k + 2\sum_{l=0}^{K} c_{kl}\theta_l + r'_k(\theta) \\
\frac{\partial^2}{\partial\theta_k\partial\theta_l}\log z(\theta) &= 2c_{kl} + r''_{kl}(\theta)
\end{aligned}$$

where r'_k and r''_{kl} are, respectively, the first partial derivative with respect to the kth component and the second partial derivative with respect to the kth and lth components of the remainder term $r(\theta)$.

The power series approximations can be equated with the expressions for the derivatives of the log normalization constants obtained previously. The remainder terms should be close to Θ for large lattices, as discussed on page 196. Thus the power series approximation to the first derivative $b_k + 2\sum_{l=0}^{K} c_{kl}\theta_l$ may be equated with the right-hand side of equation (6.46). Similarly the power series approximation to the second derivative $2c_{kl}$ can be equated with the right-hand side of equation (6.72). Hence substituting the simulation estimate γ_{kl} for the covariance $\mathrm{Cov}\,[\,V_k(\theta), V_l(\theta)\,]$ in equation (6.72) gives:

$$\gamma_{kl} - v_k v_l = 2c_{kl} + e_{kl}$$

Evaluating this expression for a particular value of θ, say θ_m, gives

$$\gamma_{mkl} - v_{mk} v_{ml} = 2c_{kl} + e_{mkl}$$

Thus using $m = 1, m \ldots, M$ will yield M different estimates of c_{kl}. These may be regularized using the techniques outlined in Section 6.4.8.

6.4.5 Improving integration

A linear or quadratic approximation to the log NC is adequate when either the lattice is large enough, or the parameter values are sufficiently close. Quadrature rules to achieve this are discussed in detail in Section 6.4.3.

A particular Binary Markov random field model, the Autologistic, has been investigated in detail by this research. For this model I have found that the assumption of linearity in the log NC is satisfactory. Any extra numerical effort is therefore unnecessary for our purposes. Results from a small experiment support this assumption that changes in the log NC are sufficiently gradual (in the parameter region of interest). As the distance between the parameters θ_A and θ_B increases however, the linearity assumption weakens. Future research could consider more computer intensive yet sophisticated methods of estimating the IMCS estimator. There are two possibilities. The first is that a more accurate surface is fitted to $\mathrm{E}_\theta\,[\,v(x)\,]$. In effect this means estimating derivatives. The second option is to use more elaborate numerical integration techniques. I briefly discuss alternatives that could be considered for both these options below.

Linear or quadratic approximations to the log NC, as functions of θ, are found to be adequate when either the lattice is large enough, or when the area of the parameter space is relatively small. Quadratic rules can provide these linear or quadratic approximations. They are discussed in detail in Section 6.4.3.

Alternatively, suppose that the lattice is not large enough to ensure that linear and quadratic approximations to the log NC are accurate. Approaches to fitting a more complex surface may be divided into parametric and nonparametric. Parametric approaches require more complex assumptions about the shape of the surface. These are difficult to apply since the behaviour of the canonical statistic is not currently well understood. Let us therefore first consider non-parametric approaches.

A non-parametric approach is more sensitive to the data in a local way. Fan & Gijbels (1996) give some modern solutions: local polynomial fitting together with derivative estimates; locally weighted smoothing; wavelet thesholding; and spline smoothing. The difficulty with all of these methods is the added complexity required to estimate the derivative

of the surface. With local polynomial fitting for instance, the first derivative is taken to be the slope of a local quadratic regression, and the second derivative is the slope of a local cubic regression.

Another example is adaptive quadrature rules (Forsythe et al. 1977), which require derivatives to be estimated to ensure the required precision of estimates. In these cases, interactive choice of evaluation points would depend on off-line simulation to compute NC ratios at specific parameter values. These evaluations could not be invoked independently of the quadrature step. For binary MRFs it is expensive to estimate v_{mk} at these parameter values, and furthermore these computations have stochastic variability.

6.4.6 Error Analysis

This method involves two stages: estimating the mean canonical statistic (MCS) for various parameter values θ_A and then integrating the MCS over combinations of the parameter values to obtain estimates of NC ratios. The two-stage method of estimation leads to two sources of error. Firstly, there is statistical error which is due to using simulations for estimating the expectation $\mathrm{E}_\theta[V(x)]$. In addition, this error needs to be adjusted if dependent simulations are used. The second source of error is numerical and is determined by the quadrature rule applied instead of analytically evaluating the integral.

Using standard sampling theory, the error introduced by the sample estimate of the expectation is straightforward to estimate. For example, with the 2-point trapezoidal rule, the variance of $\widehat{\log z_{AB}}$ is

$$\mathrm{Var}\left[\widehat{\log z_{AB}}\right] = \frac{1}{4}(\theta_{Bk} - \theta_{Ak})^2 \left(\mathrm{Var}[v_{Bk}] + \mathrm{Var}[v_{Ak}]\right).$$

Covariance terms conveniently disappear when the estimation of v_{Bk} and v_{Ak} is based on unrelated samples. The variances above may be estimated using methods commonly used for estimating variance in a MCMC chain: Integrated Autocorrelation time (IACT) (Green & Han 1990) or by approximation of the iteration sequence to a simple autoregressive series, such as an AR(1) (Tierney 1991) as discussed in Section 4.4.6. When the dependent samples are sufficiently spaced they may approximate independence so that the sample variance may even be a good estimator.

For each rule, I compare the 5-point versions for approximating the integral of the log NC ratio $\zeta_{AB} = \log z_{AB}$ to view how the factors and spacings of $\overline{V}_{Ak}$ evaluation differ. For brevity, omit the subscript k denoting the kth component of parameter θ and write θ for θ_{Ak}. Let $\delta = \theta_B - \theta_A$ measure the distance between the two parameters and $\delta_k \neq 0$ but $\delta_{k^*} = 0, k^* \neq k$. I will denote by $v(\theta)$ the simulation average of $\frac{1}{T}\sum_{t=1}^{T} V_k(x^{(t)})$ where $\{x^{(t)}\}$ are realisations from the distribution $p(x \mid \theta_m)$.

$$\begin{aligned}
\mathsf{T}_5(\zeta_{AB}) &= \tfrac{\delta}{8}\left\{v(\theta) + 2v\left(\theta + \tfrac{\delta}{4}\right) + 2v\left(\theta + \tfrac{\delta}{2}\right) + 2v\left(\theta + \tfrac{3\delta}{4}\right) + v(\theta + \delta)\right\} \\
\mathsf{R}_5(\zeta_{AB}) &= \tfrac{\delta}{5}\left\{v\left(\theta + \tfrac{\delta}{10}\right) + v\left(\theta + \tfrac{3\delta}{10}\right) + v\left(\theta + \tfrac{\delta}{2}\right) + v\left(\theta + \tfrac{7\delta}{10}\right) + v\left(\theta + \tfrac{9\delta}{10}\right)\right\} \\
\mathsf{S}_5(\zeta_{AB}) &= \tfrac{\delta}{12}\left\{v(\theta) + 4v\left(\theta + \tfrac{\delta}{4}\right) + 2v\left(\theta + \tfrac{\delta}{2}\right) + 4v\left(\theta + \tfrac{3\delta}{4}\right) + v(\theta + \delta)\right\}
\end{aligned}$$

The inaccuracy introduced by the numerical integration can be estimated since the Quadrature rules are all linear in $\overline{V}_{mk}$. Each rule can be expressed in the form

$$\mathsf{Q}_M(\zeta_{AB}) = cW^{\mathsf{T}}\overline{V}_{\cdot k} = c\sum_{m=1}^{M} w_m \overline{V}_{mk}$$

where $\mathsf{Q} \in \{\mathsf{T}, \mathsf{R}, \mathsf{S}\}$ denotes the choice of quadrature rule, W is a vector of weights governing the impact of each evaluated $\overline{V}_{mk}$; and c is a constant factor applied to these weights. Here $\overline{V}_{\cdot k}$ denotes a vector of the kth component of the mean canonical statistic evaluated at m values of θ in the interval $[\theta_A, \theta_B]$. That is

$$\overline{V}_{\cdot k} = \begin{bmatrix} \overline{V}(\theta_{1k}) \\ \overline{V}(\theta_{2k}) \\ \vdots \\ \overline{V}(\theta_{Mk}) \end{bmatrix} \quad \text{and} \quad \theta_{mk} = \theta_{Ak} + \delta_k h_m(M)$$

where $h_m(M)$ is the mth component of $h(M)$, a vector giving the spacings of θ where v is evaluated; and δ_k is a constant factor controlling the spacings.

Each of the quantities W, c and h take on different values for the different quadrature rules, as shown below.

Rule	c	$h(M)$	W
T_M	$\frac{\delta}{2(M-1)}$	$\frac{1}{M-1}\begin{bmatrix} 0 & 1 & \dots & M-1 \end{bmatrix}$	$W_i = \begin{cases} 2, & i = 2,3,\dots,M-1 \\ 1, & i = 1, M \end{cases}$
R_M	$\frac{\delta}{M}$	$\frac{1}{2M}\begin{bmatrix} 1 & 2 & \dots & M \end{bmatrix}$	$W_i = 1, \quad i = 1,2,\dots,M$
S_M	$\frac{\delta}{3(M-1)}$	$\frac{1}{M-1}\begin{bmatrix} 0 & 1 & \dots & M-1 \end{bmatrix}$	$W_i = \begin{cases} 1, & i = 1, M \\ 4, & i = 2,4,\dots,M-1 \\ 2, & i = 3,5,\dots,M-2 \end{cases}$

Note that M must be odd for the trapezoidal or Simpson's rule.

Suppose that the variances of v_{mk} are sufficiently similar and be approximated by σ_V^2, say. This occurs when the distances between θ_k values, δ, is small. Then the variances of the M-point rules are approximately

$$\begin{aligned} \text{Var}\left[\mathsf{R}_M(\zeta_{AB})\right] &\approx \frac{1}{M}\delta^2\sigma_V^2 \\ \text{Var}\left[\mathsf{T}_M(\zeta_{AB})\right] &\approx \frac{(2M-3)}{2(M-1)^2}\delta^2\sigma_V^2 \\ \text{Var}\left[\mathsf{S}_M(\zeta_{AB})\right] &\approx \frac{2(5M-6)}{9(M-1)^2}\delta^2\sigma_V^2 \end{aligned}$$

Figure 6.1 shows that using the same number of evaluation points, (*i.e.* M constant), Simpson's rule is more variable than the Trapezoidal rule, which is slightly more variable than the Rectangular rule. Even in the worst case, using only $M = 3$ points, the difference in the variability for the three rules is less than 20% of $\delta^2\sigma_V^2$.

Comparison

For our purposes, both the trapezoidal and Simpson's rules are preferable to the rectangular rule since in the latter case the function v (the IMCS estimator) is not evaluated at the 2 last points of interest. The log NC ratio would not utilize the MCMC simulations already available for the upper and lower θ values of interest.

Comparing the general expressions for these integrals shows that the main changes apart from the spacings h are the multiplicative factor c and the weights W. These have the largest impact on the variance estimates, as shown below.

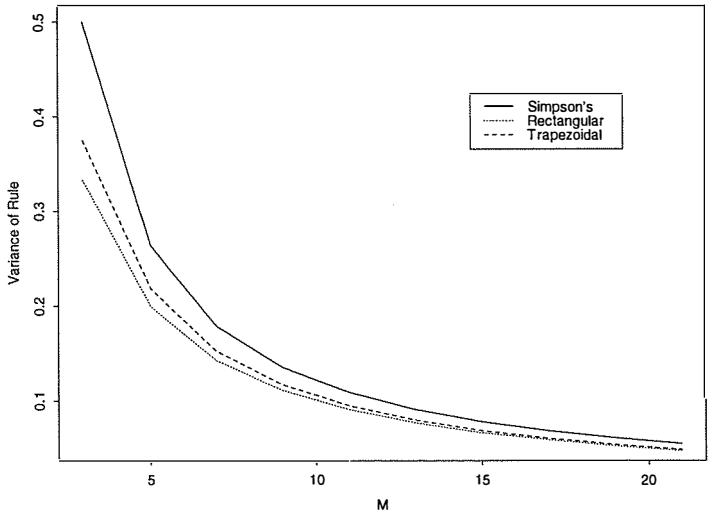


Figure 6.1: Variance of quadrature estimates of log NCR assuming equal variance at evaluation points.

To compare the inferential properties, namely bias and mean square error, of the different methods of integration, analytic expressions cannot be used or else the problem of evaluating the NC would be solved! I therefore investigate a member of the exponential family for which it is straightforward to obtain analytical results, namely independent Poisson distributions. The other reason for selecting the Poisson distribution is that the autologistic distribution reduces to a Poisson process when the dependence parameters are zero.

The joint distribution of L *iid* Poisson variates observed on the lattice $\{X_i, \quad i = 1, \ldots, L\}$, with mean intensity λ, is given by:

$$f(x \mid \lambda) = \frac{\exp\{-L\lambda\}\lambda^{\sum_{i=1}^{L} x_i}}{\prod_{i=1}^{L} x_i!}$$

or setting $\theta_O = \log \lambda$ this becomes

$$f(x \mid \lambda) = \frac{\exp\{\theta_0 \sum_{i=1}^{L} x_i - \sum_{i=1}^{L} \log x_i!\}}{\exp\{Le^{\theta_0}\}}. \tag{6.61}$$

The Poisson density can be expressed in a form similar to equation (6.2) by setting:

$$V(x) = \begin{bmatrix} \sum_{i=1}^{L} x_i \\ -\sum_{i=1}^{L} \log x_i! \end{bmatrix} \quad \text{and} \quad \theta = \begin{bmatrix} \theta_0 \\ 1 \end{bmatrix}, \quad \theta_0 = \log \lambda$$

In this Poisson case, the log NC can be obtained analytically from the denominator of equation (6.61), since it is given by

$$z(\theta) = \exp\{ne^{\theta_0}\}$$

and so

$$\log z_{AB} = n(e^{\theta_{A0}} - e^{\theta_{B0}}).$$

The first and second moments of the canonical statistic are the same since this is a Poisson distribution:

$$\begin{aligned} \mathrm{E}[V_0(x)] &= \mathrm{E}\left[\sum x_i\right] = L\lambda = Le^{\theta_0} \\ \mathrm{Var}[V_0(x)] &= \mathrm{Var}\left[\sum x_i\right] = L\lambda = Le^{\theta_0} \end{aligned}$$

The bias of the estimation of the log NCR can be computed as the difference between the expected value and the true value. The Mean Square Error (MSE) is then obtained by summing the variance and the squared bias. When computing the MSE, in practise the major contribution will be from the variance, since in this case bias can be decreased by simply increasing the notional MCMC sample size. The variance for the integration therefore varies with the squares of the weights and of the constant.

Suppose that the variance varies smoothly from one point to the next, then the (discarded) rectangular rule minimises the variance by equally weighting all points. The most variable weights arise from the use of Simpson's rule so the variance of the overall integral is expected to be higher than that for the Trapezoidal rule.

For example, a small notional simulation experiment was undertaken for each of the three integration rules, for $M = 3$, 5 or 7 points, to estimate the IMCS for the Poisson distribution with $\theta_0 = -2.2$ ($\lambda = e_0^\theta = 0.1108$) and $\theta_0 = -1.95$ ($\lambda = e_0^\theta = 0.1423$). The notional MCMC sample size was set to 10,000. Table 6.1 shows the results of this simulation experiment. I discuss the results from the first case initially, where the expected value of the IMCS estimator was 29.7.

The variances of the log NC ratios, as predicted, were smallest for the rectangular rule and then for the trapezoidal rule and they decrease as more points are evaluated to estimate the integral. They range from $\mathrm{Var}[R_7] = 1.06$ to $\mathrm{Var}[S_3] = 3.71$. The bias was 2 orders of magnitude smaller, with Simpson's rule always being the most accurate, particularly with more evaluation points. The rectangular rule was found to be only slightly less biassed than the trapezoidal rule. The MSEs were all driven by the variances since the biases were comparatively small.

6.4.7 Reparameterization

Suppose that the M-point Quadrature rule Q_M is used to compute the IMCS estimates of the NCR. Referring to equation (6.46) this will require evaluation of the mean canonical statistic v_m, and hence simulation, for the distributions $p(x \mid \theta_m)$. Each of these distributions takes on a different parameter value in the sequence $\theta_m \in [\theta_A, \theta_B]$ with $m = 1, \ldots, M$ and setting $A = 1$, $B = M$. However, this direct method of integration applies only when pairs θ_A, θ_B (and all intervening pairs) differ in precisely one component k. When the pair θ_A, θ_B differ in more than one component, then two approaches to integration may be taken: one indirect without reparameterization, and the other direct requiring reparameterization.

The first approach to integration proceeds via *indirect* integration over a series of "adjacent" pairs, such that within each pair only one component differs. Thus a "path" $\theta_A = \theta_1, \theta_2, \ldots, \theta_{M-1}, \theta_M = \theta_B$ needs to be established where consecutive pairs θ_m, θ_{m+1} are "adjacent". By definition, the log NCRs of these pairs are related via

$$z_{AB} = \frac{z(\theta_A)}{z(\theta_B)} = \frac{z(\theta_A)}{z(\theta_m)} \times \frac{z(\theta_m)}{z(\theta_B)} = z_{Am} \cdot z_{mB}$$

and therefore

$$\log z_{AB} = \log z_{Am} + \log z_{mB}. \tag{6.62}$$

Table 6.1: Comparison of Rectangular, Trapezoidal and Simpson's rule for integration of the mean canonical statistic in evaluating ratios of Normalization constants. For each rule the 3-point, 5-point and 7-point versions are considered. Two values of θ_0 (as denoted in the main column heading) are considered. The *E*xpected value, the *Bias* and *R*oot *M*ean *S*quare *E*rror of the log NC ratios is given in the columns for a notional MCMC simulation sample size of 10,000.

	$\theta_0 = -2.2$			$\theta_0 = -1.95$		
	E	**Bias**	**RMSE**	**E**	**Bias**	**RMSE**
R_3	29.731	-0.00860	1.57	38.176	-0.01105	1.78
R_5	29.737	-0.00310	1.22	38.183	-0.00398	1.38
R_7	29.738	-0.00158	1.03	38.185	-0.00203	1.17
T_3	29.779	0.03871	1.67	38.237	0.04971	1.89
T_5	29.750	0.00968	1.28	38.199	0.01243	1.44
T_7	29.744	0.00430	1.07	38.192	0.00552	1.21
S_3	29.740	0.00004	1.93	38.187	0.00005	2.18
S_5	29.740	0.00000	1.40	38.187	0.00000	1.59
S_7	29.740	0.00000	1.15	38.187	0.00000	1.31

The alternative approach to integration is a *direct* approach. It is possible to transform the parameter space of any dimension so that any *pair* of θ values no longer differ in more than one component. Suppose that a pair of θ values, (θ_A, θ_B) can be transformed to (ϕ_A, ϕ_B) via the linear translation

$$\theta = G\phi \quad \text{and} \quad \phi = H\theta \quad \text{where} \quad G = H^{-1} \tag{6.63}$$

Here the rotation H is chosen so that the resulting coordinates ϕ differ in only one component. In our example situation, this requires

$$\begin{aligned} \phi_{Aj} &= \phi_{Bj} \qquad \forall j \neq k \\ \phi_{Aj} &\neq \phi_{Bj} \qquad j = k. \end{aligned}$$

The density $p(x \mid \theta)$ can be rewritten in terms of ϕ as

$$p(x \mid \theta) = \frac{\exp\{\phi^{\mathsf{T}} W(x)\}}{z(\phi)} \tag{6.64}$$

where $W(x) = G^{\mathsf{T}} V(x)$. The normalization constant is unchanged although it can now be re-expressed in terms of ϕ

$$z(\phi) = \sum_{x \in \Omega} \exp\{\phi^{\mathsf{T}} W(x)\} = z(\theta).$$

Thus the NC ratio $z_{AB} = \frac{z(\theta_A)}{z(\theta_B)}$ is equivalent to $\frac{z(\phi_A)}{z(\phi_B)}$. By symmetry the formulation of the IMCS estimator based on ϕ is the same as that based on θ:

$$\log z_{AB} = \int_{\phi_B}^{\phi_A} \mathrm{E}_\phi [\, W_k(x) \,]\, d\phi \tag{6.65}$$

This integral can be approximated by an M-point quadrature rule Q_M on the interval $[\phi_{Ak}, \phi_{Bk}]$ based on evaluations of $\overline{W}_k$,

$$\widehat{\log z_{AB}} \approx \mathrm{Q}_M(\phi_{Ak}, \phi_{Bk}, \overline{W}_k)$$

where

$$\begin{aligned} \phi_{Ak} &= H_{k\cdot}^{\mathsf{T}} \theta_{Ak} \\ \phi_{Bk} &= H_{k\cdot}^{\mathsf{T}} \theta_{Bk} \\ \overline{W}_k &= G_{\cdot k}^{\mathsf{T}} v_k \quad . \end{aligned}$$

Here the notation $H_{k\cdot}$ denotes the kth row of H and $G_{\cdot k}$ denotes the kth column of G.

To illustrate these two approaches let us consider an example situation where three dimensional parameters θ_A and θ_B differ in components $k = 1, 2$. Writing the origin as $A = (\theta_0^*, \theta_1^*, \theta_2^*)$, then the other point B can be expressed in relation to A as $B = (\theta_0^*, \theta_1^* + \delta_1, \theta_2^* + \delta_2)$. The intermediary points $C = (\theta_0^*, \theta_1^*, \theta_2^* + \delta_2)$ and $D = (\theta_0^*, \theta_1^* + \delta_1, \theta_2^*)$ differ in only one component from both A and B. The relationship between the four points A, B, C, and D is illustrated in Figure 6.2.

The first *indirect* approach may proceed by consideration of either series of pairs $(\theta_A, \theta_C), (\theta_C, \theta_B)$ or $(\theta_A, \theta_D), (\theta_D, \theta_B)$. Within each pair only one component differs, either $k = 1$ or $k = 2$.

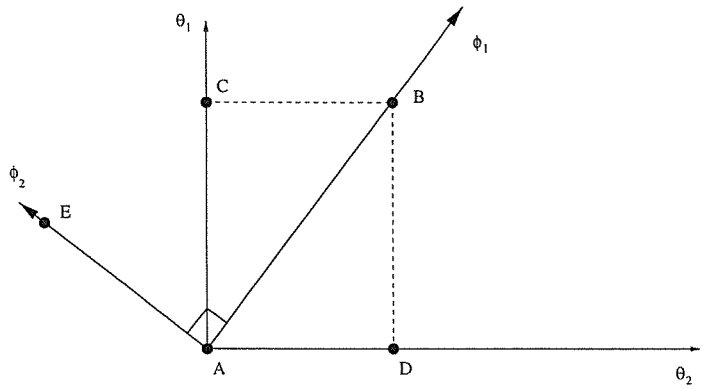


Figure 6.2: The relationship between a pair of parameter values θ_A and θ_B which differ in 2 dimensions. Two paths via the θ_C or θ_D are available which enforce changes in only one dimension at a time.

If the 2-point Trapezoidal rule is used then the log NCR can be approximated using relationship equation (6.62)

$$\log z_{AB} \approx -\tfrac{1}{2}\{\delta_2(\overline{V}_{A2} + \overline{V}_{C2}) + \delta_1(\overline{V}_{C1} + \overline{V}_{B1})\} \tag{6.66}$$

Alternatively the *direct* approach requires translation from (θ_1, θ_2) space to (ϕ_1, ϕ_2) space such that ϕ_1 is equal for points A and B. As shown in Figure 6.2, the line $\overleftrightarrow{AB}$ connecting A and B has equation

$$\theta_1 = \frac{\delta_1}{\delta_2}\theta_2 + \theta_1^* - \frac{\delta_1}{\delta_2}\theta_2^*.$$

Setting

$$\phi_1 = -\theta_1 + \frac{\delta_1}{\delta_2}\theta_2 + \theta_1^* - \frac{\delta_1}{\delta_2}\theta_2^*$$

ensures that ϕ_1 is constant for points A and B.

The line $\overleftrightarrow{AE}$ which is perpendicular to $\overleftrightarrow{AB}$ and intersects this line at point A has equation

$$\theta_1 = -\frac{\delta_2}{\delta_1}\theta_2 + \theta_1^* + \frac{\delta_2}{\delta_1}\theta_2^*.$$

Setting

$$\phi_2 = -\theta_1 - \frac{\delta_2}{\delta_1}\theta_2 + \theta_1^* + \frac{\delta_2}{\delta_1}\theta_2^*$$

gives a line perpendicular to $\overleftrightarrow{AB}$. A transformation mapping θ onto ϕ such that at least one component of ϕ is constant is defined above using ϕ_1 and ϕ_2. This can be written

$$\begin{bmatrix} \phi_1 \\ \phi_2 \end{bmatrix} = \begin{bmatrix} -\frac{\delta_1}{\delta_2}\theta_2^* + \theta_1^* \\ \frac{\delta_2}{\delta_1}\theta_2^* + \theta_1^* \end{bmatrix} + \begin{bmatrix} -1 & \frac{\delta_1}{\delta_2} \\ -1 & -\frac{\delta_2}{\delta_1} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$$

which is in the form $\phi = H_0 + H\theta$. Hence after mapping A and B from θ-space onto ϕ-space, using $[H_0 \mid H]$, their coordinates are

$$\begin{aligned} A_\phi &= (0, 0) \\ B_\phi &= (0, -\frac{1}{\delta_1}(\delta_1^2 + \delta_2^2)) \end{aligned}$$

Applying equation (6.63) requires solution of

$$\phi^{\mathsf{T}} W(x) \equiv \theta^{\mathsf{T}} V(x) = (H^{-1}(\phi - H_0))^{\mathsf{T}} V(x).$$

Now

$$H^{-1} = \frac{\delta_1 \delta 2}{\delta_1^2 + \delta_2^2} \begin{bmatrix} -\frac{\delta_2}{\delta_1} & -\frac{\delta_1}{\delta_2} \\ 1 & -1 \end{bmatrix}$$

and

$$(H^{-1} H_0)^{\mathsf{T}} = -(\theta^*)^{\mathsf{T}} = -(\theta_1^* \theta_2^*)$$

These results give

$$W(x) = \left(H^{-1} - \begin{bmatrix} -\theta_1^* \phi_1 & 0 \\ 0 & \theta_2^* \phi_2 \end{bmatrix} \right)^{\mathsf{T}} V(x)$$

Substituting into equation (6.65) followed by some arithmetic gives

$$\widehat{\log z_{AB}} \approx \tfrac{1}{2} (\phi_{A2} - \phi_{B2}) \left(\overline{W}(\phi_{A2}) + \overline{W}(\phi_{B2}) \right) \quad (6.67)$$

$$= -\tfrac{1}{2} [\delta_1 (v_{A1} + v_{B1}) + \delta_2 (v_{A2} + v_{B2})] . \quad (6.68)$$

Comparing equation (6.68) to equation (6.66) shows that the only difference is that v_{A1} replaces v_{C1}. Essentially this means that for the *direct* approach, in each dimension the "closer" point has been used in the numerical integration.

6.4.8 Regularization of pairwise IMCS estimates

Each path is comprised of consecutive sub-paths, each involving integration between pairs of points in θ space. Using a numerical integration method such as quadrature ensures that this integration comprises linear combinations of the expected mean canonical statistic corresponding to these pairs of θ values. Essentially although the NC ratios may theoretically be evaluated by different paths through parameter space, this gives rise to a number of different estimates, due to the differences in statistical precision involved. For instance, paths may be more precise if they minimize the number of nodes and the length of paths containing changes in a particular component of θ. This section discusses averaging paths to obtain an overall estimate of the log NC, balancing the best and worst paths. In practice, this requires estimation of many paths, and is therefore computationally expensive. An alternative, presented in Section 6.4.9, shows how to find a single optimal path. The question of which path was not practical for the *dingo* case study, since a sensitivity analysis with respect to paths showed that there was little important difference between log NC values computed. However in other applications the choice of path may be important.

The results from evaluating the NC ratios via different 'paths' through the parameter values gives rise to a regularization problem. Although each of the pairwise log NC ratios are based on the same underlying NCs, they have all been estimated separately. This leads to an accumulation of numerical and statistical errors imposed by estimating ratios pairwise. The pairwise estimates can be combined in a linear regression to retrieve the underlying normalizing constants as coefficients, assuming that the NCs combine linearly (according to the design matrix) to form the integrals. This is equivalent to constructing an averaged path estimate. Expressed in matrix form, the underlying relationship to the NCs can be

written as follows:

$$\begin{bmatrix} \log z_{AB} \\ \log z_{AC} \\ \log z_{AD} \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & 0 & \dots & 0 \\ 1 & 0 & -1 & 0 & \dots & 0 \\ 1 & 0 & 0 & -1 & \dots & 0 \\ \vdots & & & & & \vdots \end{bmatrix} \begin{bmatrix} \zeta_A \\ \zeta_B \\ \zeta_C \\ \vdots \end{bmatrix}$$

or $\log z = A\zeta$.

This reduces to a regression problem with known design matrix A and observations $\log z$. The unknown parameters to be estimated are $\{\zeta_m\}$. As for the situation with estimation of fixed effects in a linear model, each ζ_m parameter can only be estimated relative to the others, and one linear constraint needs to be chosen. Here the corner-point constraint is used and we set $\zeta_A = 0$. If one component ζ_A is set to be zero, then all results will be relative to point A. Another alternative is the sum-to-zero constraint. The variances of each measurement can be approximated as shown below. The estimates $\widehat{\zeta}$ can then be obtained using least squares estimates with known variance in a standard statistical package.

Estimates of a pairwise log NCR discussed in this section are based on two steps: numerical (quadrature) integration; and the simulation average of the mean canonical statistic. Both steps are based on linear combinations. The quadrature step is based on linear combinations of the mean canonical statistics, which themselves are linear combinations of simulated canonical statistics. The variance of the resulting log NCR is therefore a linear combination of the variances of the simulation averages v_{mk} and the covariances between components of the mean canonical statistic estimates from the same model m, v_{mk} and v_{mj}. For example, in the illustration above, the 2-point trapezoidal rule is used to estimate the log NCR between θ_A and θ_C which differ in two dimensions via reparameterization:

$$\mathsf{T}_2(\log z_{AB}) = -\tfrac{1}{2}\left[\delta_1(\overline{V}_{A1} + \overline{V}_{B1}) + \delta_2(\overline{V}_{A2} + \overline{V}_{B2})\right]. \tag{6.69}$$

The variance of this estimate is

$$\begin{aligned} \mathrm{Var}\left[\,\mathsf{T}_2(\log z_{AB})\,\right] \;=\; & \tfrac{1}{4}\left\{\delta_1^2\left(\mathrm{Var}\left[\,\overline{V}_{A1}\,\right] + \mathrm{Var}\left[\,\overline{V}_{B1}\,\right]\right)\right. \\ & +\delta_2^2\left(\mathrm{Var}\left[\,\overline{V}_{A2}\,\right] + \mathrm{Var}\left[\,\overline{V}_{B2}\,\right]\right) \\ & + 2\delta_1\delta_2\left(\mathrm{Cov}\left[\,\overline{V}_{A1}, \overline{V}_{A2}\,\right]\right. \\ & \left.\left. + \mathrm{Cov}\left[\,\overline{V}_{B1}, \overline{V}_{B2}\,\right]\right)\right\} \end{aligned} \tag{6.70}$$

The variance of the estimate obtained via integration over an indirect path of adjacent θ values may be computed in a similar manner.

When θ_A and θ_B differ in three dimensions it is possible though tedious to derive expressions similar to equation (6.69) and equation (6.70).

6.4.9 Optimal path

Gelman & Meng (1998, §4) conduct a theoretical investigation of the Monte Carlo quadrature estimate $\widehat{\lambda}_g$ by focussing on its variance. For our exponential family model, they show that the optimal choice for the Monte Carlo importance sampling function g is given by

$$\begin{aligned} g(\theta) \;&\propto\; \mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log h(x \mid \theta)\right)^2\right]^{\frac{1}{2}} \\ &=\; \left\{\mathrm{E}_\theta\left[\,V^2(x)\,\right]\right\}^{\frac{1}{2}} \end{aligned}$$

where the expectation is with respect to the conditional distribution of X given θ. Alternatively one could suggest a one-to-one increasing transformation ψ of θ, with $g(\theta) = \psi$, so that the resulting optimal density $p(\psi)$ is uniform. This gives

$$h(x \mid \psi) = \exp\{g^{-1}(\psi)^{\mathsf{T}} V(x)\}.$$

Now

$$\begin{aligned} p(\psi) &\propto \left\{ \mathrm{E}_\psi \left[\left(\frac{\partial}{\partial \psi} \log h(x \mid \psi) \right)^2 \right] \right\}^{\frac{1}{2}} \\ &= \left\{ \mathrm{E}_\psi \left[\left(\frac{d\theta}{d\psi} \theta \frac{\partial}{\partial \theta} \log h(x \mid \theta) \right)^2 \right] \right\}^{\frac{1}{2}} \\ &= \left\{ \frac{1}{g'(\theta)^2} \mathrm{E}_\theta \left[V(x)^2 \right] \right\}^{\frac{1}{2}}. \end{aligned}$$

So in summary, requiring $p(\psi) \propto$ constant gives

$$g(\theta) \propto \int \sqrt{\mathrm{E}_\theta \left[V(x)^2 \right]} d\theta.$$

In order to compute $\widehat{\lambda}_g$ using this method requires estimation of $\mathrm{E}_\theta \left[V(x)^2 \right]$ and estimation of $g(\theta)$. This would suggest use of a two-stage approach where $\mathrm{E}_\theta \left[V(x)^2 \right]$ is first estimated over an equally spaced grid of θ values; $g(\theta)$ is then estimated and a new grid equally spaced in $g(\theta)$ can be used for the next iteration.

For example, in the *dingo* case study, interest is focussed on θ values in a three-dimensional cube. A suitable grid for θ is:

$$\{(\theta_0, \theta_1, \theta_2) : \theta_0 \in \{-1.7, -1.8, \ldots, -2.2\},\ \theta_1, \theta_2 \in \{0.00, 0.25, \ldots, 1.50\}\}. \tag{6.71}$$

The design of the grid is justified in more detail in Section 6.5. Interpretation of how these parameter values affect the likelihood is more easily achieved using the local conditional likelihood of equation (6.74). Comparison of estimates of the log NC computed between any pair of θ points in the Θ cube given in equation (6.71) showed less than a 5% change depending on the path taken. For instance, two choices of the integration path taken from $\theta = (-2.2, 0, 0)$ to $\theta = (-1.7, 1, 1)$ are: via $(-2.2, 1, 0)$ and $(-2.2, 1, 1)$; or via $(-1.7, 0, 0)$ and $(-1.7, 1, 0)$.

A not uncommon situation would be that the variability of each component of θ is different. Heuristically, the optimal path would therefore minimize the number of changes for those components of θ with the highest variance. An expression to define the optimal path for integration was given in (Gelman & Meng 1998, p168, near equation (18)), however this is inaccurate.

6.4.10 Higher order derivatives

Contrast the system of $K+1$ equations given in equation (6.46) and equation (6.47) for the first order partial derivatives of $\log z(\theta)$ with the higher-order derivatives of $\log z(\theta)$. The second order partial derivative with respect to the kth and lth components of θ is equivalent

to the covariance between these two components:

$$\begin{aligned}\frac{\partial^2}{\partial\theta_k\partial\theta_l}\log z(\theta) &= \frac{\partial}{\partial\theta_k}\sum_{x\in\Omega}\frac{V_l(x)\exp\{\theta^{\mathsf{T}}V(x)\}}{z(\theta)} \\ &= \mathrm{E}_\theta\left[V_k(x)V_l(x)\right] - \mathrm{E}_\theta\left[V_k(x)\right]\mathrm{E}_\theta\left[V_l(x)\right] \\ &= \mathrm{Cov}\left[V_k(x), V_l(x)\right] \qquad (6.72)\end{aligned}$$

The third order partial derivative with respect to the jth, kth and lth components of θ is the third order cumulant:

$$\begin{aligned}\frac{\partial^3}{\partial\theta_j\partial\theta_k\partial\theta_l}\log z(\theta) &= \frac{\partial}{\partial\theta_j}\sum_{x\in\Omega}\frac{V_k(x)V_l(x)\exp\{\theta^{\mathsf{T}}V(x)\}}{z(\theta)} \\ &\quad -\frac{\partial}{\partial\theta_j}\sum_{x\in\Omega}\frac{V_k(x)\exp\{\theta^{\mathsf{T}}V(x)\}}{z(\theta)}\sum_{x\in\Omega}\frac{V_l(x)\exp\{\theta^{\mathsf{T}}V(x)\}}{z(\theta)} \\ &= \mathrm{E}_\theta\left[V_j(x)V_k(x)V_l(x)\right] - \mathrm{E}_\theta\left[V_j(x)\right]\mathrm{E}_\theta\left[V_k(x)V_l(x)\right] \\ &\quad -\mathrm{E}_\theta\left[V_k(x)\right]\mathrm{E}_\theta\left[V_j(x)V_l(x)\right] - \mathrm{E}_\theta\left[V_l(x)\right]\mathrm{E}_\theta\left[V_j(x)V_k(x)\right] \\ &\quad +2\mathrm{E}_\theta\left[V_j(x)\right]\mathrm{E}_\theta\left[V_k(x)\right]\mathrm{E}_\theta\left[V_l(x)\right] \\ &= \kappa_3(V_j(x), V_k(x), V_l(x)) \qquad (6.73)\end{aligned}$$

For exponential family distributions having parameter θ of dimension higher than four, expressions for high order partial derivatives can be obtained in the same manner. However the number of terms quickly escalates, and the simplicity of the first order derivatives is lost.

I now explore the nature of the higher order derivatives of $\log z(\theta)$. As shown by Equations (6.72) and (6.73) the higher order derivatives of $\log z(\theta)$ are in fact equivalent to cumulants of $V(x)$ with respect to the density $p(x)$ given in equation (6.2). Cumulants $\{\kappa_j\}$ can alternatively be obtained from the terms of the log moment generating function since $\log m(t) = \sum_{j=0}^{\infty}\kappa_j\frac{t^j}{j!}$ where the moment generating function $m(t)$ evaluates to:

$$m(t) = \mathrm{E}_\theta\left[e^{t^{\mathsf{T}}V(x)}\right] = \frac{\sum_{x\in\Omega}e^{t^{\mathsf{T}}V(x)}e^{\theta^{\mathsf{T}}V(x)}}{z(\theta)} = \frac{z(\theta+t)}{z(\theta)}.$$

This relationship between the moment generating function and the NC ratio was noted in Geyer & Thompson (1992) as discussed in Section 6.2.5.

However a simplification is available since the canonical statistics V are approximately Normally distributed particularly for lower values of spatial dependence, as the lattice size increases, $L \to \infty$, due to the Central Limit theorem. The simplification results since all cumulants other than the first and second cumulants are zero for the Normal distribution. Hence the third and higher order partial derivatives of the logged normalization constant will be zero for large lattices. So the closer the higher derivatives will be to 0, the closer the joint distribution of the canonical statistics $(V_0, V_1, V_2, \ldots)$ will be to the multivariate Normal distribution. In addition, the central limit theorem implies that the larger the lattice then the closer the distribution of the sums comprising V will be to the normal distribution. Thus the quadratic expansion about θ should be a reasonable approximation for large lattices. For a small lattice, a locally linear/quadratic approximation may be more appropriate.

6.5 Case study

6.5.1 Design

For this simulation case study, the selected design parallels that investigated in more detail in Pettitt & Low Choy (1999), and later in Chapter 7. These studies analyze results of a field experiment investigating how well various chemicals attract dingos, the Australian native dog. Binary visit/no-visit observations were obtained over a two-dimensional grid. One dimension represents the spatial placement of chemical pairs along a (road) transect. The other dimension was due to the observations being repeated over seven days, and thus represents the temporal displacement of the data. See Pettitt & Low Choy (1999) for a detailed discussion of the experimental design.

In the first analysis the dependence between observations was captured in a parameter p_i, the probability of dingo presence for (spatio-temporal) site i. Depending on model choices, this probability could vary discretely over blocks or smoothly over adjacent sites. Within a linear model framework with treatment effects being treated as fixed effects, analysis proceeded via a frequentist approach utilizing the EM algorithm of Dempster et al. (1977) to obtain MLEs. In the second analysis we introduce an underlying process for the presence/absence of dingos over the grid, and incorporate this layer into a hierarchical model. Analysis of this hierarchical model proceeds via Bayesian methods utilizing MCMC techniques to obtain posterior distributions of parameters. In this case study however we ignore the overlying process which models the treatment effects of chemical attractiveness. Instead we focus on the underlying two-dimensional process describing the presence and absence of dingos at sites throughout the lattice. Estimation of the Normalization Constant for the presence/absence process is required for a fully Bayesian approach to inference, and has thus motivated this chapter.

The essential characteristics of the dingo data to be investigated here mirror the spatio-temporal presence/absence process used in the Bayesian approach in Chapter 7. These characteristics cover the extent of the grid and type of response variable; the prevalence parameter; and the degree of dependence.

Extent

The data are binary with 0 and 1 indicating absence and presence respectively. The grid is two-dimensional and medium-sized with approximately 1000 sites: $L = 135$(horizontal) $\times$ 7(vertical) = 945 sites. The perimeter to area ratio of the study area is more than double that of a square grid with the same number of sites. Edge effects will therefore have greater impact on analysis.

Prevalence

The overall prevalence of dingos or probability of presence is fairly low and previous studies (Pettitt & Low Choy 1999) its value is in the range 0.10–0.15.

It is definitely *not* expected that presence and absence of dingos is evenly balanced. The number of sites having the value 0 should be different to those that have the value 1. This will usually be the case for practical applications in spatial statistics of the AL(3) model.

Grimmett (1973) and others (Strauss 1975) show that the joint distribution has an equivalent formulation as a conditional distribution as a Markov Random Field. This conditional form describes the probability of presence at a particular site which depends *only*

on the presence/absence at neighbouring sites, and on the parameter θ.

$$p(x_i|x_{-i};\theta) = \frac{\exp\{x_i(\theta_0 + \sum_{k=1}^{K}\theta_k(x_{i:+k} + x_{i:-k}))\}}{1 + \exp\{\theta_0 + \sum_{k=1}^{K}\theta_k(x_{i:+k} + x_{i:-k})\}} \tag{6.74}$$

For simplicity we assume a first order neighbourhood, so that $K = 2$. Let π_{hv} be the probability of presence given h horizontal neighbours of site i being present, and v vertical neighbours of site i being present. Then applying equation (6.74) gives:

$$\begin{aligned}\pi_{hv} &= p(x_i = 1|x_{i:+1} + x_{i:-1} = h; x_{i:+2} + x_{i:-2} = v; \theta) \\ &= \frac{\exp\{\theta_0 + h\theta_1 + v\theta_2\}}{1 + \exp\{\theta_0 + h\theta_1 + v\theta_2\}}\end{aligned}$$

and therefore

$$\theta_0 + h\theta_1 + v\theta_2 = \log\left(\frac{\pi_{hv}}{1 - \pi_{hv}}\right) = \text{logit}\,(\pi_{hv}) \tag{6.75}$$

We may apply this result to the situation where no neighbours are present, *i.e.* $h = 0$ and $v = 0$. Then

$$\theta_0 = \text{logit}\,(\pi_{00})$$

Hence given an estimate of π_{00}—the probability of a presence given no neighbouring presences—we can 'select' an appropriate value for θ_0.

In this study since we are interested in prevalences being in the range 0.10–0.15, this corresponds to θ_0 values given below:

Pr{presence | no first-order neighbours} π_{00}	θ_0
0.10	-2.1972
0.125	-1.9459
0.15	-1.7346

Thus θ_0 values chosen for case study are: -2.2, -1.95, -1.7. We make the simplifying assumption that maximum probability of presence occurs when all neighbours are present. We make a further simplifying assumption that all dependence parameters $\{\theta_k\}$ are positive, that is, the more neighbours are present, the higher the probability of presence. These assumptions reflect the situation encountered in Pettitt & Low Choy (1999) and Chapter 7. If some dependence parameters are negative then we could instead aim for maximum probability of presence under different neighbourhood conditions, *e.g.* no neighbours present.

Degree of dependence

Applying the result in equation (6.75) to the case where all first order neighbours are present, *i.e.* $h = 2$ and $v = 2$, we obtain an expression for the combined dependence in both dimensions:

$$\theta_{\text{tot}} = \theta_1 + \theta_2 = \frac{1}{2}(\text{logit}\,(\pi_{22}) - \theta_0)$$

	Pr{presence|all first order neighbours present} $\pi_{22} = p(x_i = 1 \mid h_i = 2, v_i = 2)$		
θ_{tot}	$\theta_0 = -1.7$	$\theta_0 = -1.95$	$\theta_0 = -2.2$
0	0.154	0.512	0.858
0.5	0.125	0.450	0.964
1	0.100	0.786	0.955
1.5	0.332	0.741	0.943
2	0.279	0.690	0.987
2.5	0.231	0.909	0.983
3	0.574	0.886	0.978

Finally the asymmetry of dependence between the two dimensions can be described by the arithmetic difference

$$\begin{aligned} \theta_{\text{diff}} = \theta_1 - \theta_2 &= \text{logit}\,(\pi_{10}) - \text{logit}\,(\pi_{01}) \\ \text{or} &= \frac{1}{2}\left[\text{logit}\,(\pi_{20}) - \text{logit}\,(\pi_{02})\right] \end{aligned} \tag{6.76}$$

The values of θ chosen for the experiment reflect those chosen in Chapter 7. For the case where $\theta_1, \theta_2 > 0$, the probability of dingo presence will take on its highest value when all neighbours are present, *i.e.* for maximum π_{22}. The lowest value corresponds to π_{00}. The middle value we have chosen is π_{11} for situations where one each of the horizontal and vertical neighbours are present. The high, middle, and low probabilities associated with these values are tabulated below in Table 6.2.

6.5.2 Base models

The extent of models to be investigated in this case study has now been determined in the discussion above. Recall however that it is the *Ratio* of Normalization Constants that is of interest. A comparison of the estimation methods (Importance Sampling MC, MCMC Ratio, RLR and IMCS) is therefore best facilitated by the selection of a few base models with parameters θ_B, so that the ratios $\frac{z(\theta_A)}{z(\theta_B)}$ may be compared for the 48 values of θ_A. These 48 values were chosen to coincide with those from the first simulation experiment presented in Chapter 7, and correspond to a full $3 \times 4 \times 4$ factorial design on

$$\begin{aligned} \theta_0 &\in \{-1.7, -1.95, -2.2\} \\ \theta_1 &\in \{0, 0.5, 1.0, 1.5\} \\ \theta_2 &\in \{0, 0.5, 1.0, 1.5\} \end{aligned} \tag{6.77}$$

Two base models were chosen for investigation. The first θ_B corresponds to the zero dependence situation and low prevalence. The second situation is a model centrally located in the parameter space, and so θ_C has slightly lower prevalence but medium levels of dependence in both directions.

model index	**base model**	corresponding θ
$m = 1$	B	$\theta_B = (-1.7, 0.0, 0.0)$
$m = 17$	C	$\theta_C = (-1.95, 0.5, 0.5)$

To evaluate path sampling and Reverse Logistic Regression it is necessary to compare results for a broader lattice to those obtained from a finer lattice. For the broad lattice

θ_0 prevalence	θ_2 (vertical)	θ_1 (horizontal) 0	0.5	1	1.5
-1.7	0	0.154	0.154	0.154	0.154
		0.154	0.231	0.332	0.45
		0.154	0.332	0.574	0.786
	0.5	0.154	0.154	0.154	0.154
		0.231	0.332	0.45	0.574
		0.332	0.574	0.786	0.909
	1	0.154	0.154	0.154	0.154
		0.332	0.45	0.574	0.69
		0.574	0.786	0.909	0.964
	1.5	0.154	0.154	0.154	0.154
		0.45	0.574	0.69	0.786
		0.786	0.909	0.964	0.987
-1.95	0	0.125	0.125	0.125	0.125
		0.125	0.19	0.279	0.389
		0.125	0.279	0.512	0.741
	0.5	0.125	0.125	0.125	0.125
		0.19	0.279	0.389	0.512
		0.279	0.512	0.741	0.886
	1	0.125	0.125	0.125	0.125
		0.279	0.389	0.512	0.634
		0.512	0.741	0.886	0.955
	1.5	0.125	0.125	0.125	0.125
		0.389	0.512	0.634	0.741
		0.741	0.886	0.955	0.983
-2.2	0	0.100	0.100	0.100	0.100
		0.100	0.154	0.231	0.332
		0.100	0.231	0.450	0.690
	0.5	0.100	0.100	0.100	0.100
		0.154	0.231	0.332	0.450
		0.231	0.450	0.690	0.858
	1	0.100	0.100	0.100	0.100
		0.231	0.332	0.450	0.574
		0.450	0.690	0.858	0.943
	1.5	0.100	0.100	0.100	0.100
		0.332	0.450	0.574	0.690
		0.690	0.858	0.943	0.978

Table 6.2: Triplets of low, middle and high conditional probabilities of presence (assuming positive dependence parameters) at a single site are given for various levels of prevalence (θ_0) and horizontal (θ_1) and vertical (θ_2) dependence.

we have selected some points from the design space described in equation (6.77) to yield a minimal $2 \times 2 \times 2$ lattice:

$$\begin{aligned} \theta_0 &\in \{-1.7, -2.2\} \\ \theta_1 &\in \{0, 1.0\} \\ \theta_2 &\in \{0, 1.0\} \end{aligned} \tag{6.78}$$

For the finer lattice we have selected an additional point in each dimension located between the extremes given in equation (6.78) to yield a $3 \times 3 \times 3$ lattice:

$$\begin{aligned} \theta_0 &\in \{-1.7, -1.95, -2.2\} \\ \theta_1 &\in \{0, 0.5, 1.0\} \\ \theta_2 &\in \{0, 0.5, 1.0\} \end{aligned} \tag{6.79}$$

The endpoints selected for the path are $\theta_A = (-1.7, 0, 0)$ and $\theta_B = (-2.2, 1, 1)$. This allows us to investigate changing from a model with no dependence (random) but higher prevalence to a model with high dependence (clustered) and lower prevalence. It is of interest how selection of the path is affected by the properties of the prevalence and dependence parameters.

6.5.3 Simulation Study Results

In this section, estimates of log NCs are reported for the case study detailed earlier in Sections 6.5.1 and 6.5.2. Four methods are investigated: the importance sampling method using the importance sampling function discussed in Section 6.2.1 (ISMC); dependent Monte Carlo of Section 6.2.5 (MCMC); the Reverse Logistic regression of Section 6.3 (RLR) and the Integrated Mean Canonical Statistical method (IMCS), a special case of the Path sampling method of Section 6.4.

Recall that the models represented in the denominators of the NC ratios were defined with model B being zero spatio-temporal dependence, and model C being a model with medium spatio-temporal dependence. The model indexed by $m = 1$ was denoted by B and had parameter $\theta = c(-1.70, 0.0, 0.0)$. Model $m = 17$ was denoted by C and had parameter $\theta = c(-1.95, 0.5, 0.5)$.

With the IMCS method, the quadrature rule used is trapezoidal with order 2. The representative path was generally selected to first traverse θ_0, then θ_1 and θ_2 in that order. Informal results showed that there was little difference in the order that the components of θ were traversed in constructing the path. Complete results are shown in Table 6.3 for comparison of the IMCS, importance sampling and dependent Monte Carlo estimators of the log NC ratios.

Standard errors are not reported, since these were generally very large for the Monte Carlo approaches (generally of the same order as the estimated log NC ratio), and less than 5% for the IMCS approach.

Table 6.3: Results from estimation of log NC ratios using various methods, namely, Integrated Mean Canonical Statistic (IMCS), Importance Sampling Monte Carlo and Markov Chain Monte Carlo.

base model	**model**			**Log NC ratio estimates**		
θ_m	θ_A			$\lambda(A,m) = \log c(\theta_A)/c(\theta_m)$		
m	θ_{A0}	θ_{A1}	θ_{A2}	IMCS	IS MC	MC MC
C	-1.7	0	0	10.06	11.25	11.00
C	-1.7	0	0.5	23.77	24.40	24.29
C	-1.7	0	1	50.53	42.64	50.87
C	-1.7	0	1.5	121.43	62.53	81.35
C	-1.7	0.5	0	26.21	26.68	26.76
C	-1.7	0.5	0.5	46.65	42.02	46.30
C	-1.7	0.5	1	94.95	62.24	75.72
C	-1.7	0.5	1.5	261.20	83.49	106.61
C	-1.7	1	0	59.34	48.55	53.77
C	-1.7	1	0.5	104.1	66.06	77.54
C	-1.7	1	1	236.58	79.12	107.62
C	-1.7	1	1.5	495.47	96.51	138.90
C	-1.7	1.5	0	149.03	68.23	87.13
C	-1.7	1.5	0.5	329.36	81.68	110.36
C	-1.7	1.5	1	601.42	102.32	139.96
C	-1.7	1.5	1.5	914.90	120.94	171.36
C	-1.95	0	0	-22.92	-21.54	-22.77

continued on next page

Table 6.3: (continued from previous page)

base model	model			Log NC estimates		
θ_m	θ_A			$\lambda(A,m) = \log c(\theta_A)/c(\theta_m)$		
m	θ_{A0}	θ_{A1}	θ_{A2}	IMCS	IS MC	MC MC
C	-1.95	0	0.5	-14.00	-13.01	-13.38
C	-1.95	0	1	3.14	3.65	3.07
C	-1.95	0	1.5	47.29	23.01	32.12
C	-1.95	0.5	0	-12.42	-11.25	-11.75
C	-1.95	0.5	0.5	0.00	0.00	0.00
C	-1.95	0.5	1	27.00	20.88	23.38
C	-1.95	0.5	1.5	127.97	38.26	53.72
C	-1.95	1	0	8.62	5.61	7.60
C	-1.95	1	0.5	31.91	21.99	27.06
C	-1.95	1	1	104.98	38.11	54.37
C	-1.95	1	1.5	352.90	52.64	85.62
C	-1.95	1.5	0	69.35	30.88	39.23
C	-1.95	1.5	0.5	173.96	42.97	61.72
C	-1.95	1.5	1	384.32	56.76	87.00
C	-1.95	1.5	1.5	660.90	74.58	118.11
C	-2.2	0	0	-49.43	-48.85	-52.14
C	-2.2	0	0.5	-43.67	-42.90	-43.74
C	-2.2	0	1	-32.82	-33.41	-32.93
C	-2.2	0	1.5	-6.44	-16.57	-14.12
C	-2.2	0.5	0	-42.66	-41.73	-42.75
C	-2.2	0.5	0.5	-35.10	-34.47	-34.97
C	-2.2	0.5	1	-20.02	-22.12	-20.67
C	-2.2	0.5	1.5	30.24	-3.57	6.29
C	-2.2	1	0	-29.39	-30.08	-29.72
C	-2.2	1	0.5	-17.46	-19.84	-18.30
C	-2.2	1	1	15.03	-2.55	3.99
C	-2.2	1	1.5	214.73	13.78	32.85
C	-2.2	1.5	0	12.79	-8.91	-6.54
C	-2.2	1.5	0.5	50.63	0.49	14.02
C	-2.2	1.5	1	193.08	16.61	37.21
C	-2.2	1.5	1.5	451.92	37.64	65.03
B	-1.7	0	0	0.00	0.00	0.00
B	-1.7	0	0.5	13.86	13.15	13.28
B	-1.7	0	1	46.11	31.39	32.99
B	-1.7	0	1.5	134.64	51.28	53.85
B	-1.7	0.5	0	16.37	15.42	15.21
B	-1.7	0.5	0.5	37.17	30.77	30.07
B	-1.7	0.5	1	98.63	50.99	49.64
B	-1.7	0.5	1.5	316.68	72.23	70.58
B	-1.7	1	0	57.06	37.30	36.03
B	-1.7	1	0.5	101.16	54.81	50.46
B	-1.7	1	1	282.56	67.87	67.06

continued on next page

Table 6.3: (continued from previous page)

base model θ_m	model θ_A			Log NC estimates $\lambda(A,m) = \log c(\theta_A)/c(\theta_m)$		
m	θ_{A0}	θ_{A1}	θ_{A2}	IMCS	IS MC	MC MC
B	-1.7	1	1.5	584.92	85.26	87.43
B	-1.7	1.5	0	185.01	56.98	57.76
B	-1.7	1.5	0.5	319.22	70.43	72.00
B	-1.7	1.5	1	586.17	91.07	87.40
B	-1.7	1.5	1.5	848.87	109.68	104.67
B	-1.95	0	0	-32.98	-32.79	-33.02
B	-1.95	0	0.5	-24.05	-24.26	-24.17
B	-1.95	0	1	-3.96	-7.60	-9.22
B	-1.95	0	1.5	51.09	11.76	9.87
B	-1.95	0.5	0	-22.47	-22.51	-22.69
B	-1.95	0.5	0.5	-10.06	-11.25	-11.30
B	-1.95	0.5	1	23.44	9.62	5.81
B	-1.95	0.5	1.5	157.35	27.01	26.47
B	-1.95	1	0	2.67	-5.65	-4.28
B	-1.95	1	0.5	25.75	10.73	9.54
B	-1.95	1	1	130.12	26.86	24.71
B	-1.95	1	1.5	456.12	41.38	43.44
B	-1.95	1.5	0	80.03	19.63	17.18
B	-1.95	1.5	0.5	154.91	31.72	31.46
B	-1.95	1.5	1	415.18	45.51	46.47
B	-1.95	1.5	1.5	682.09	63.33	61.88
B	-2.2	0	0	-60.10	-60.10	-61.64
B	-2.2	0	0.5	-54.44	-54.15	-54.27
B	-2.2	0	1	-42.06	-44.66	-43.51
B	-2.2	0	1.5	-9.44	-27.82	-27.34
B	-2.2	0.5	0	-53.47	-52.98	-53.70
B	-2.2	0.5	0.5	-46.09	-45.72	-46.21
B	-2.2	0.5	1	-27.87	-33.37	-33.59
B	-2.2	0.5	1.5	37.71	-14.82	-16.37
B	-2.2	1	0	-38.26	-41.33	-40.00
B	-2.2	1	0.5	-26.21	-31.10	-29.77
B	-2.2	1	1	-18.00	-13.81	-15.34
B	-2.2	1	1.5	296.95	2.53	0.43
B	-2.2	1.5	0	6.01	-20.17	-21.16
B	-2.2	1.5	0.5	38.54	-10.76	-8.32
B	-2.2	1.5	1	257.42	5.36	6.61
B	-2.2	1.5	1.5	557.71	26.39	21.63

Reverse logistic regression results are reported in Table 6.4. This method was much more computationally intensive, requiring weeks rather than days of computing time. In addition, it was often difficult to obtain convergence, due to problems with local minima or very flat likelihood surfaces. The number of sweeps in the table attests to difficulties with achieving convergence, which occurred for higher levels of dependence. Results were

not obtainable due to these problems for some parameter sets, which are omitted from the table. Two different 'powerdowns' were used representing the baseline $a = 1$ and squaring the likelihood $a = 2$. These results show that no benefit is obtained from powering up the likelihood in most cases, although for some cases the maximization algorithm converged more rapidly when $a = 2$.

Table 6.4: Simulation study results for Reverse Logistic Regression

Models compared						Log NC ratio estimates			
base model θ_B			model θ_A			Power factor $a = 1$		Power factor $a = 2$	
θ_{B0}	θ_{B1}	θ_{B2}	θ_{A0}	θ_{A1}	θ_{A2}	$\log \widehat{z}_{AB}$	N_{sweeps}	$\log \widehat{z}_{AB}$	N_{sweeps}
-1.7	0	0	-1.95	0	0	-32.8813	113	-32.8812	47
-1.7	0	0	-2.20	0	0	-59.2853	139	-59.2893	>640
-1.7	0	0	-1.70	0.5	0	15.6816	77	15.6814	>240
-1.7	0	0	-1.95	0.5	0	-22.7730	34	-22.7729	46
-1.7	0	0	-2.20	0.5	0	-52.8441	61	-52.8450	73
-1.7	0	0	-1.70	1.0	0	49.7252	9	49.7252	9
-1.7	0	0	-1.95	1.0	0	-2.8413	50	-2.8397	26
-1.7	0	0	-2.20	1.0	0	-40.5652	128	-40.5643	> 4000
-1.7	0	0	-1.70	0	0.5	13.3551	90		
-1.7	0	0	-1.95	0	0.5	-24.2464	33		
-1.7	0	0	-2.20	0	0.5	-53.7522	75		
-1.7	0	0	-1.70	0.5	0.5	35.3481	117	26.1901	92
-1.7	0	0	-1.95	0.5	0.5	-10.9246	86	-14.0849	49
-1.7	0	0	-2.20	0.5	0.5			-45.4861	>3840

6.5.4 Discussion

Since the IMCS method of estimation is of central interest, only some relationships between the methods are discussed. Initially we report the relationship found between the two Monte Carlo methods: Importance sampling Monte Carlo (ISMC) and Markov Chain Monte Carlo (MCMC). Since these two methods appear to produce similar results, only one is used for comparison to IMCS. Finally RLR results are more comparable to IMCS, so this comparison is given last.

Relationship between Importance Sampling Monte Carlo (ISMC) and Dependent Monte Carlo (MCMC)

We first review the theoretical differences between ISMC and MCMC estimation approaches, as presented in 6.2.1 and 6.2.5 respectively.

ISMC requires independent simulation of two groups of variates $\{x_A^{(t)}, x_B^{(t)}\}$ IID Bernoulli $t = 1, 2, \ldots, T$. Since simulations are from the grossly oversimplified version of the probability distribution of interest T must be large to compensate. The exponentiated contribution $h(x|\theta)$ to the likelihood is then arithmetically averaged over each set A and B separateley

to provide the numerator and denominator of the estimator which, as discussed previously, are both subject to numerical instability.

MCMC requires dependent simulations from just the baseline model $x_B^{(t)} \sim p(x|\theta_B), \quad t = 1, \ldots, T$. These simulations are more computationally involved compared to the simple IID Bernoulli variables of ISMC, but an advantage is that they are from the probability distribution of interst. The dependent nature of the chain will require greater length.

The exponential contribution is adjusted for the numerator model $\frac{h(x|\theta_A)}{h(x|\theta_B)}$ and then arithmetically averaged over the set B simulations. Although this computation is more centred it still suffers from the same numerical instability as ISMC.

Thus on application, the relative efficacy of ISMC and MCMC estimators will be driven by the following determinants:

computational resources The number of simulations that will affect the "representativeness" of the ISMC estimator, and the convergence of simulations to the correct probability distribution in the MCMC approach.

choice of models The relative "distance" between models A and B will affect the size of the log likelihood contributions $h(\cdot)$ as well as the relative magnitude of the numerator and denominator in the ISMC estimator and the scaling effect on the summand in the MCMC estimator.

degree of dependence Models with low spatial dependence will be well approximated by the Bernoulli simulations, making ISMC and MCMC simulated variates similar in distribution. In addition the variability of these estimators is much lower for low spatial dependence. With higher spatial dependence simulations for the MCMC approach will take longer to converge.

sampling variation This will of course contribute to observed differences between estimators.

size of the lattice As the lattice increases in size, so too does inherent variability in variates generated for the MCMC method.

An empirical comparison of results can be achieved via simple linear regression. We have plotted values of the log NC ratio, ISMC and MCMC estimates given in Table 6.3, in Figure 6.3, including the line of equality. These figures contain results for comparison to base model B where $m = 1$, and were reproduced for comparison to the other base model considered, C, where $m = 17$.

From Figure 6.3, it is evident that these two methods provide estimates of the log NC ratios that are highly comparable. The values appear to have almost perfect correspondence. The sum against difference of estimates plot also shows a more or less random error pattern, and the frequency plot of the differences is reasonably Normal in shape. It is interesting that samples from independent identical prior distributions, the basis of ISMC, could produce such similar results as the MCMC method. Clearly, the numerical feature of summing exponentiated sums of presences are dominating both these methods. Figure 6.4 shows diagnostics arising from a regression of MCMC on ISMC estimates. Clearly this relationship is very strong, providing further evidence of the link between the two methods. In Figure 6.5 models deviating from a perfectly linear relationship between ISMC and MCMC results are characterised by the total amount of spatio-temporal dependence, $\theta_{A,\text{tot}} = \theta_{A1} + \theta_{A2}$. The plot of MCMC *vs* ISMC estimates shows that the models that deviate from the linear

relationship mostly have higher values of dependence. A sum and difference version of Figure 6.5 better highlights the large deviations from the straight line fit. The random nature of Figure 6.6 demonstrates that the precise linear fit is consistent across all models.

When the baseline model changes from B to C, the spatial dependence in the baseline model also changes from zero to a medium level. Figure 6.7 shows that the relationship between the ISMC and MCMC estimates of the log NC ratio is very similar using a baseline of model C compared to results based on model C. A slight difference is noted in the influential and outlier points. Models 16 and 12, with high prevalence $\theta_0 = -1.7$, high temporal dependence $\theta_{A2} = 1.5$ and medium to high spatial dependence $\theta_{A2} \in \{1.0, 1.5\}$, are highly influential in the linear fit for both baselines. Model 32, with lower prevalence $\theta_0 = -1.95$ but the same spatio-temporal dependence as model 16, is only highly influential in the linear fit when baseline model C is used. Model 15, with the same θ as model 16 except for slightly lower temporal dependence, is only influential in the linear fit when baseline model B is used. These highly influential points (identified by large Cook's distance) strongly affect the overall view of the consistency between the ISMC and MCMC log NC ratio estimates.

Outlier models identified when the C baseline is used correspond to $\theta \in \{(-1.7, 1, 0), (-1.95, 1, 1.5), (-1.95, 1.5, 1.5)\}$. When the B baseline is used, outlier models correspond to $\theta \in \{(-2.2, 1.5, 1.5), (-1.7, 0, 1.5), (-1.7, 1.5, 1.5)\}$. Thus we may conclude that, not surprisingly, using a baseline model with θ situated at one corner of Θ space means that the most inconsistency between log NC ratio estimates occurs for the model at the opposite corner of Θ space. In addition, given either a baseline model with θ situated at one corner or in the centre of the parameter space, the MCMC and ISMC estimates of log NC ratios are inconsistent for models with the same prevalence level but very high temporal dependence.

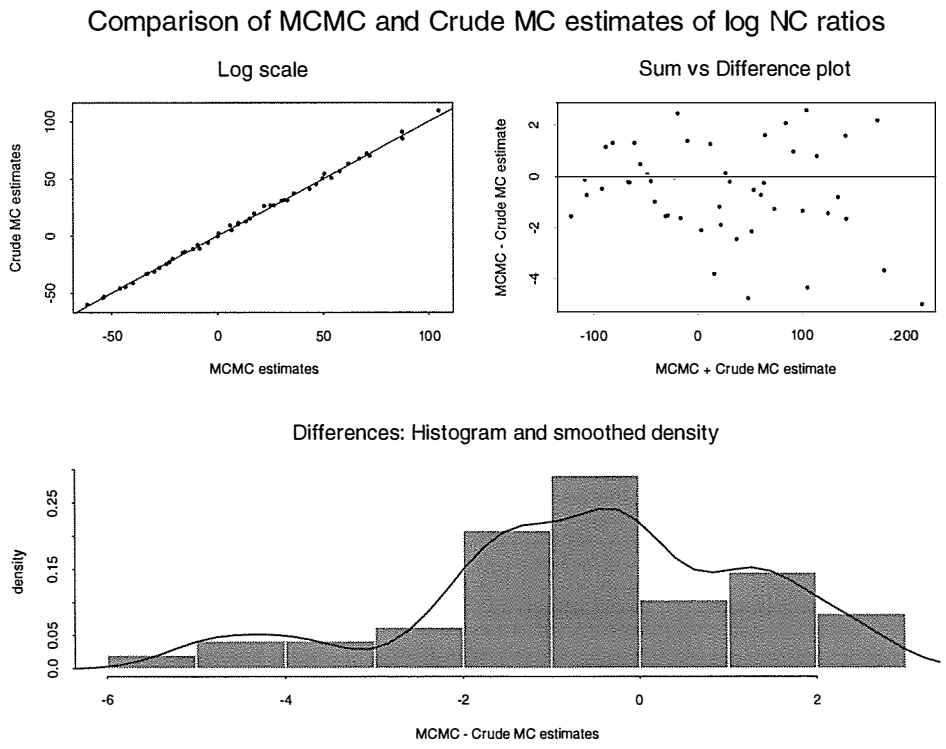


Figure 6.3: Log NC ratios compared to base model B obtained via Importance Sampling MC *vs* MCMC Ratio. Baseline model is B, $m = 1$

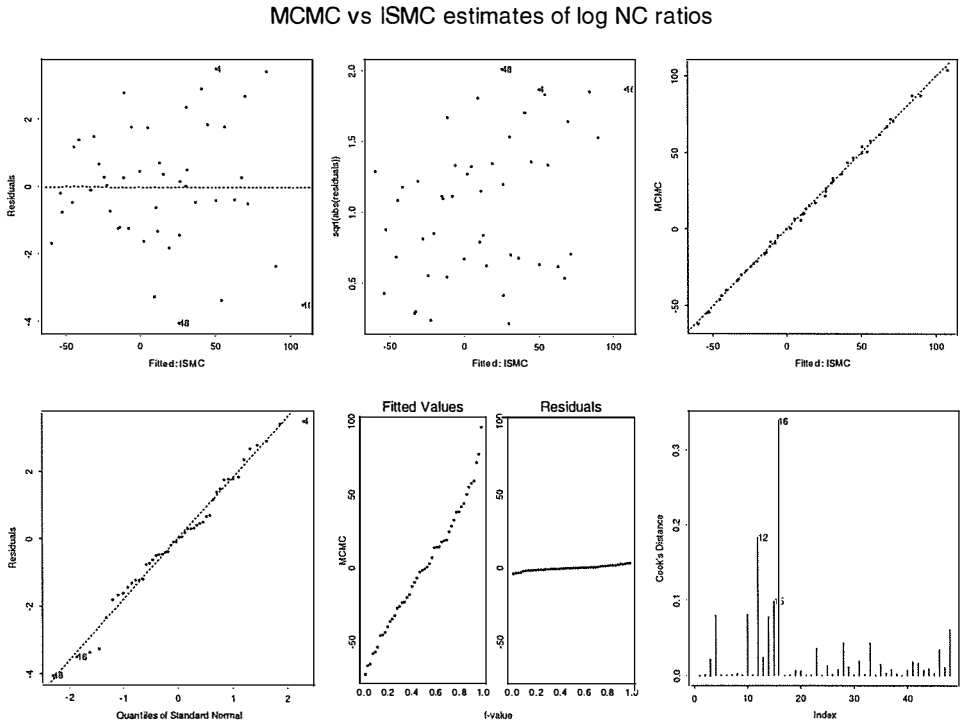


Figure 6.4: Model diagnostics for fitting linear relationship between ISMC and MCMC Ratio Estimates. Baseline model is B, $m = 1$

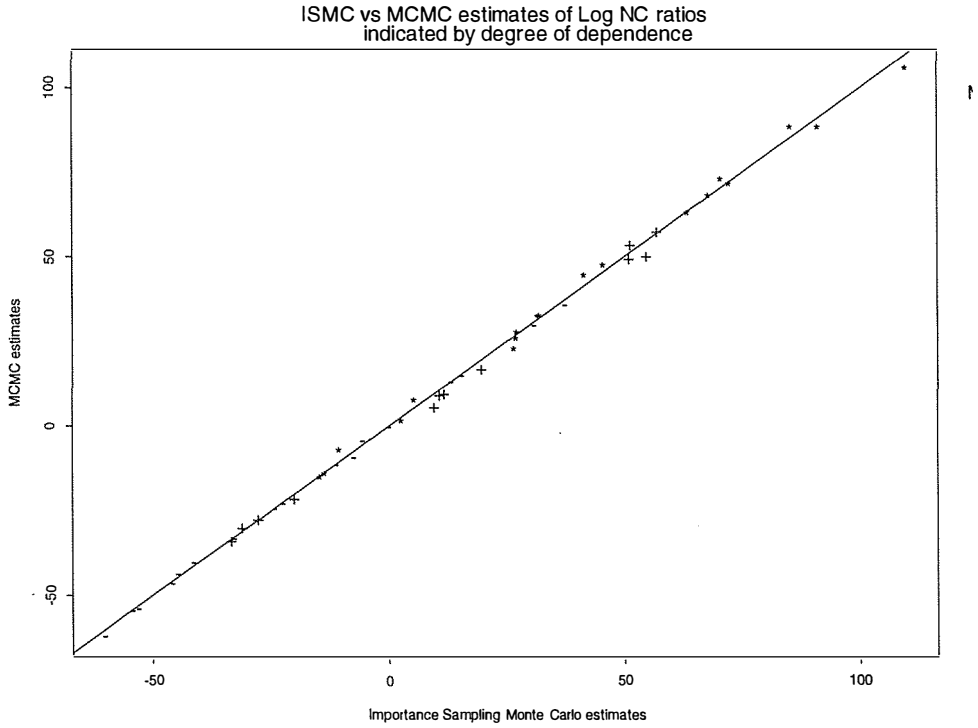


Figure 6.5: Comparison of log NC ratios obtained via MCMC Ratio and Importance Sampling Monte Carlo (ISMC) with different levels of dependence $\theta_{A,\text{tot}}$ annotated on the plot. $-$ denotes low, $+$ moderate, and $*$ high values of $\theta_{A,\text{tot}}$. Baseline model is B, $m = 1$

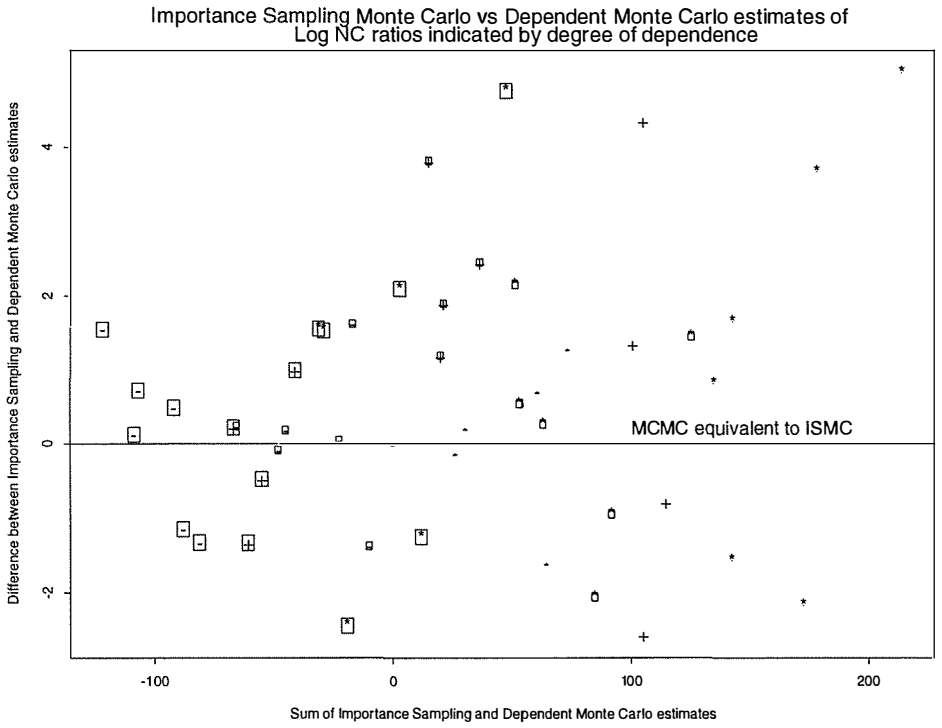


Figure 6.6: As for Figure 6.5 except that (1) this is a sum and difference plot; (2) the size of the squares is inversely proportional to θ_0 (*i.e.* large squares are for $\theta_0 = -2.2$, medium for -1.95 and small for -1.7.) Baseline model is B, $m = 1$

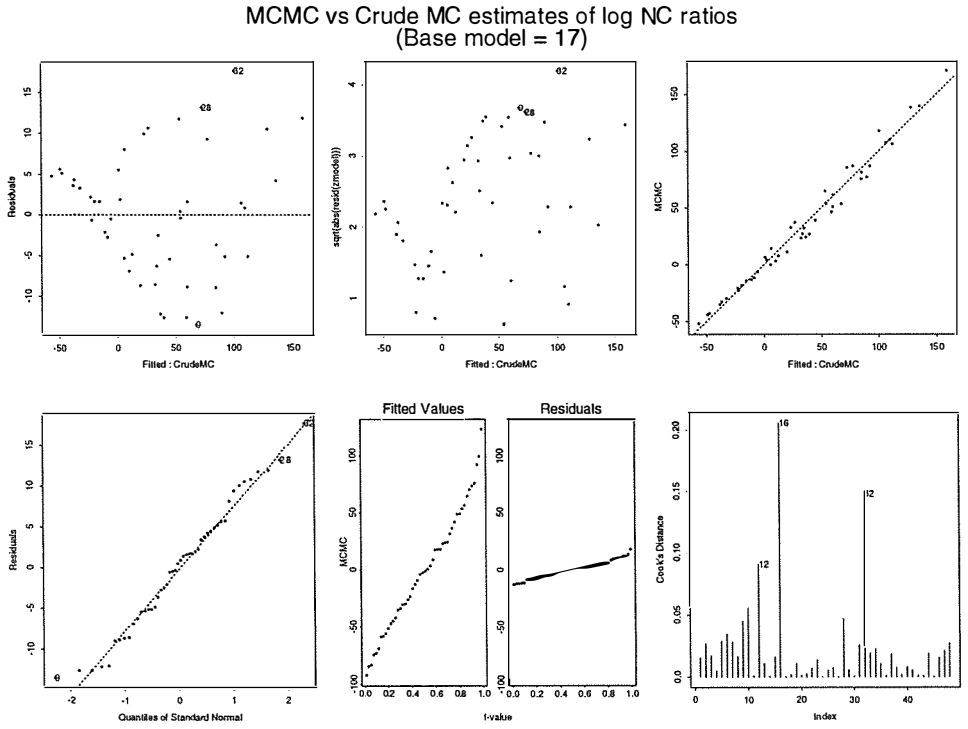


Figure 6.7: Model diagnostics for fitting linear relationship between ISMC and MCMC Ratio Estimates. Baseline model is C, $m = 17$

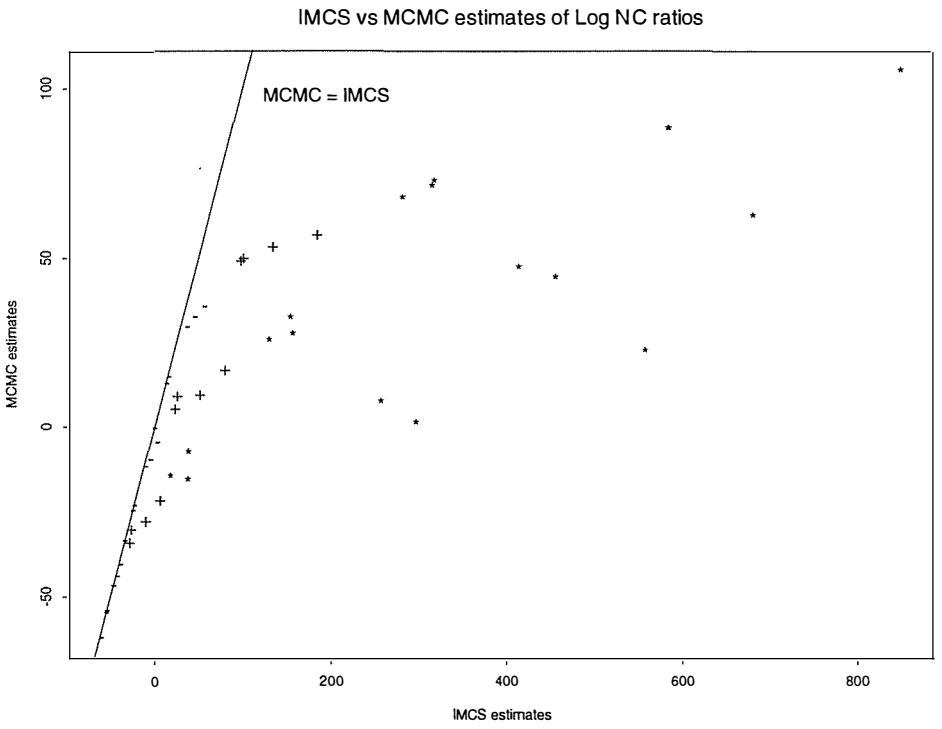


Figure 6.8: Comparison of log NC ratios obtained via MCMC Ratio and IMCS with different levels of dependence θ_{tot} annotated on the plot. $-$ denotes low, $+$ moderate, and $*$ high values of θ_{tot}. Baseline model is B, $m = 1$.

Relationship between Integrated Mean Canonical Statistic (IMCS) and Dependent Monte Carlo (MCMC)

From the analysis so far, it appears that ISMC and MCMC give results which are consistent with one another. However, if we inspect Table 6.3 we note that there are inconsistencies between the values of IMCS and the values of ISMC and MCMC, and this relationship is seen in the plot of values of ISMC against MCMC in Figure 6.8. In this figure we have divided the model parameter values into three groups corresponding to low, medium or high spatial dependence ($\theta_{A,tot} = \theta_{A1} + \theta_{A2}$) in the model being compared to the baseline model B. From Figure 6.8, we also see that the discrepancies tend to occur for the high spatial dependence values, suggesting a failure in the MCMC method for these parameter values.

Since the ISMC estimates and the Dependent MCMC ratio estimates are so similar, and the ISMC estimates have such a large variance, only MCMC is compared to IMCS. Alternatively I could have compared the average of MCMC and ISMC to IMCS.

Qualitatively, it is possible to gain some understanding of the relationship between IMCS, and the pair MCMC and ISMC. Note that IMCS requires simulation from *all* models, $X^{(t)} \sim f(x|\theta_m)$; $m = 1, \ldots, 48$. ISMC needs just one simulation $X^{(t)} \sim Bernoulli(p; L)$ and then the different NC ratio estimators can be computed for different θ_m. MCMC requires just one simulation per base model $X^{(t)} \sim f(x|\theta_m)$; $m = B, C$.

The ISMC and MCMC Ratio estimators both suffer from the same computational problem. The summation required to calculate the point estimate is a sum of *exponentiated*

terms. These terms $\{V(x^{(t)})\}$ varied from nearly zero to a few hundred, on the logarithmic scale, for parameters θ_B and θ_C.

Although the central limit theorem (for ISMC) and the Ergodic Theorem (for MCMC Ratio) ensure consistency of estimates, a very large number of samples are required in practice for the point estimate to be accurate.

This is supported by comparison of the variances for all the estimators which shows that ISMC and the MCMC Ratio estimators have variances much larger than those for IMCS.

Reverse Logistic Regression vs Integrated Mean Canonical Statistic

Reverse Logistic Regression and the Integrated Mean Canonical Statistic method have close correspondence on estimates of log NC ratios, for those parameter values with low spatio-temporal dependence, as demonstrated in Table 6.5. For models with high spatio-temporal dependence, it proved difficult to obtain numerical convergence of results for the RLR method, and so in these cases IMCS is the outright winner.

RLR however is very expensive and requires simulation from all models and also from each mixture of models for which pairwise NC ratios are to be computed. This can range from $48 \times 47 \div 2 = 1128$ pairwise mixture models each taking about a day's computations, to a single mixture model with 48 components taking more than 1128 days' computation!

From Table 6.5 it is evident that the MCMC estimates of log NC ratios are underestimating by larger amounts as spatio-temporal dependence increases. Inspection of Table 6.3 shows that the models with highest differences in log NC ratios are those with the highest levels of dependence.

Table 6.5: Comparison of Reverse Logistic Regression and Integrated Mean Canonical Statistic. Base model $\theta_A = (-1.7, 0, 0)$

Model compared θ_A			**Log NC ratio estimates**	
			RLR[1]	**IMCS**
θ_{A0}	θ_{A1}	θ_{A2}	$\log \widehat{z}_{AB}$	$\log \widehat{z}_{AB}$
-1.95	0	0	-32.8813	-32.98
-2.20	0	0	-59.2853	-60.10
-1.70	0.5	0	15.6816	16.37
-1.95	0.5	0	-22.7730	-22.47
-2.20	0.5	0	-52.8441	-53.47
-1.70	1.0	0	49.7252	-57.06
-1.95	1.0	0	-2.8413	2.67
-2.20	1.0	0	-40.5652	-38.26
-1.70	0	0.5	13.3551	13.86
-1.95	0	0.5	-24.2464	-24.05
-2.20	0	0.5	-53.7522	54.44
-1.70	0.5	0.5	35.3481	37.17
-1.95	0.5	0.5	-10.9246	10.06
-2.20	0.5	0.5		-46.09
-1.75	1.0	0.5	122.0513	101.16

6.6 Conclusions

These results are very promising and show that it is indeed possible to obtain estimates of the log NC ratios using the method of IMCS for the autologistic model. These results are applicable to the wider class of Exponential Family models including MRF models such as the auto-Binomial and the auto-Poisson. The adjective "intractable" now no longer applies to the estimation of Normalization constant ratios for the Ising/autologistic model. Estimation of log NC ratios is now feasible, but still requires intensive computation and sensible searching strategies to make the most of the off-line estimation of log NC ratios required for each parameter combination.

Computation of the log NC ratios allows hierarchical models based on these models to be implemented without approximating dependent likelihood by an independent likelihood as is current in the literature (Preisler 1993, Heikkinen & Högmander 1994, Högmander & Møller 1995, Denham & Mengersen 1999). These log NC ratios are harnessed in Chapter 7 for inference in a fully Bayesian approach to analysis of extensions to the hierarchical model given in Chapter 5.

The most successful method for computing log NC ratios appears to be the IMCS method. When compared to other Monte Carlo methods, both importance sampling and dependent sampling, it is far more accurate, particularly for situations with high spatio-temporal dependence, where approximate inference methods fall down. In comparison to the Reverse Logistic Regression method (Geyer 1996), IMCS is comparable for low spatio-temporal dependence situations and is feasible to compute for high spatio-temporal dependence situations, where it appears very difficult to obtain results for RLR.

Essentially, the IMCS technique is implemented by discretising the parameter space and then for each parameter vlaue, simulating values from the autologistic distribution so that values of means of canonical statistics can be estimated. Then log normalization ratios are calculated by combining possible paths through the parameter space. For the three-dimensional regular rectangular lattice considered here we have only considered paths which change one dimension at a time. There is further work to do, as Gelman & Meng (1998) suggest, on how to best combine all possible paths between two parameter values on the lattice. Here we have considered only a single representative path for each estimate.

The IMCS method can obviously be improved by any improvement to MCMC methods which lead to more efficient estimation of the canonical statistics. Such a method might be perfect sampling, as identified in Chapter 4.

Chapter 7

Extended Hierarchical Bayesian Model for 2D binary lattice data with underlying spatial dependence

Contents

7.1 Introduction 219
7.2 Extension I 223
7.2.1 Data model for y 223
Dingo case study 223
Intermediate step towards generalization 224
Generalized Linear Model for y 224
7.2.2 Underlying spatio-temporal dependence z 225
7.2.3 Joint distribution 226
7.2.4 Prior for spatio-temporal dependence θ 226
7.2.5 Hyperparameters ϕ 227
7.2.6 Prior for coefficients of covariates, α 227
7.2.7 Posterior distributions 228
7.2.8 Marginal Posterior distribution for α 228
7.2.9 Marginal Posterior distribution for z_{st} 229
7.2.10 Marginal Posterior distribution for θ 230
7.2.11 Computational issues: MCMC design 230
The q_k parameters 230
The unknown z_{st} parameters 230
The θ parameters 231
7.3 Extension II 231
7.3.1 Data model for y 232
General formulation for y_{st} 232
7.3.2 Joint data model for y 232
Properties of the probability transition matrix 233
7.3.3 Underlying presence/absence model Z 234
7.3.4 Prior for parameters of Presence/absence θ 234

7.3.5 Hyperprior parameters ϕ . . . 235
7.3.6 Prior for chemical attractiveness parameters α_k . . . 235
7.3.7 Prior for time dependence parameter β . . . 235
7.3.8 Marginal Posterior distributions . . . 236
Marginal Posterior for q_k . . . 236
Marginal Posterior for unknown Z_{st} . . . 239
Marginal Posterior for θ . . . 241
Marginal Posterior for β . . . 242
7.3.9 Computation via MCMC . . . 242
7.3.10 MCMC sampler for β . . . 242
MCMC sampler for α_k . . . 243
MCMC sampler for Z_{st} . . . 243
7.3.11 MCMC sampler for θ . . . 243
7.4 Results: overview . . . 243
7.5 Results: EXTENSION I, Experiment 1 . . . 244
7.5.1 Posterior distribution of θ . . . 244
Marginal posterior distributions of each θ component . . . 245
Conditional Posterior distributions . . . 245
Pairwise Posterior distributions . . . 245
Posterior distribution of three-dimensional θ . . . 250
7.5.2 Proposal distribution of θ . . . 250
7.5.3 Posterior distribution of natural statistics of presence/absence . . . 255
7.5.4 Posterior distribution of effects of explanatory variables . . . 255
7.6 Results: EXTENSION I, Experiment 2 . . . 262
7.6.1 Posterior distribution of θ . . . 262
7.6.2 Posterior distribution of natural statistics of presence/absence . . . 263
7.6.3 Posterior distribution of effects of explanatory variables . . . 264
7.7 Results: EXTENSION II, Experiment 1 . . . 265
7.7.1 Posterior distribution of α, β . . . 265
7.7.2 Posterior distribution of dingo presence . . . 268
7.7.3 Posterior distribution of θ . . . 269
7.8 Discussion . . . 271

7.1 Introduction

Time present and time past
Are both perhaps present in time future,
And time future contained in time past.
- *T. S. Elliott* Four Quartets *'Burnt Norton' (1936).*

In this thesis the basic problem considered has been how to model a binary bivariate spatio-temporal process. Success and failure at a particular site-time combination not only depends on explanatory variables but also on underlying presence patterns in space and time. Approaches considered previously in this thesis are: hierarchical and frequentist in Chapter 3; and hierarchical and Bayesian in Chapter 5. Henceforth these approaches are referred to as the basic Bayesian and frequentist models respectively. Each modelling approach considered thus far has been aimed at improving our knowledge and understanding of the problem.

Alternatives to the basic Bayesian and frequentist models have already been discussed in Section 2.4 and other parts of the thesis. In particular, some other methods which might find application to this situation are: aggregation over space or time (Section 2.4.2); flat modelling approaches rather than hierarchical approaches (page 26); frequentist rather than Bayesian approaches (Chapter 3). We discard the first method of aggregation since, for the applications of interest, it is important to investigate the effect of both space and time simulataneously. We adopt a hierarchic view rather than a flat view as it allows separate modelling of error distributions due to different processes: at the data level or at the underlying spatio-temporal process level. In addition, the resulting hierarchy divides the analytical and computational problems into smaller and more manageable problems. Finally, modelling approaches investigated fall into both the frequentist and Bayesian camps.

As discussed in more detail in Section 3.4, the main drawback of the approach taken with the basic frequentist model was that the underlying spatio-temporal process needed to be treated in an *ad hoc* fashion due to the constraints (conceptual and computational) imposed by the frequentist approach, rather than as an integral part of the model. Furthermore, although point estimates and standard errors were obtained using maximum likelihood, these are only two-number summaries based on the likelihood of the test statistics given the data.

A Bayesian approach (basic Bayesian model) was therefore considered to permit a more thorough exploration of the distributions of the parameters and to take advantage of MCMC methods applied to hierarchical Bayesian models to fully integrate the underlying spatio-temporal process. This inital model comprised one upper and one lower layer. The data layer described $p(y|z,q)$ the probability of success given presence and explanatory variables (equation (5.1)). Another layer described $p(z|\theta)$ the underlying spatio-temporal patterns in presence given overall prevalence and spatial and temporal dependence parameters (equations (5.3) and (5.4)). Parameters governing the underlying spatio-temporal process had to be fixed for analysis. This problem was specifically addressed by two methods: consideration of Bayes Factors for comparing variations to the basic Bayesian model; and modelling extensions to the basic Bayesian model in EXTENSIONS I AND II, which were only made possible by theoretical investigations made in Chapter 6.

This chapter extends the hierarchical Bayesian modelling approach (basic Bayesian model) presented in Chapter 5. Two extensions to this initial model are considered. The first allows more flexibility in modelling underlying spatio-temporal presence. The depth of the hierarchy is extended via the addition of another layer $p(\theta)$ to capture the randomness

in prevalence and spatial and temporal dependence parameters. We call this adaptation of the model EXTENSION I, and examine it in detail in Section 7.2.

The second extension investigates flexibility in modelling time dependence. In the initial model, the effect of time is modelled by a temporal dependence parameter θ_2 incorporated in the underlying spatio-temporal dependence level. An alternative is to model the effect of time at the upper level in the data model. In effect this "flattens" the complete model by combining the error due to time dependence with the residual error after accounting for the effect of the covariates on the observed data. Considering this extension allows us to investigate whether the separation of contributions to error distributions in EXTENSION I affects results. This extension is examined in detail in Section 7.3 and labelled EXTENSION II.

We explore the modelling theory for each of the two extensions in a way that permits easy comparison. To achieve this we harness DAGs, a powerful graphical tool introduced in Appendix Chapter A for displaying modelled relationships between variables. DAGs are introduced in Chapter A as a means of illustrating relationships between entities, including random parameters, known data and logical relationships. In Figure 5.1 the original Bayesian hierarchical model was presented and provides the basis for the two extensions considered here. Each of EXTENSIONS I AND II is illustrated by graphical representations in Figures 7.1 and 7.2 respectively. Using this visual representation of the models, it is easy to see that the only differences between the models are: whether θ is considered fixed (basic Bayesian model) or random (EXTENSIONS I AND II)—as shown by square rather than rounded boxes; whether θ has two or three components; whether β and y_{t-1} enter into the right hand branch of the model for q link function. The interpretation of these two different methods of modelling temporal effects can be contrasted by considering how time, likelihood of success, and likelihood of presence are handled.

The common modelling theory supporting inference is then explored in more detail for each of the extended models later in Sections 7.2 and 7.3. These two sections detail theory underlying the estimation of posterior distributions of the parameters and is presented in a similar format for each extension. Modelling simulations are designed for exploring the parameter spaces of and for deriving the posterior distributions of parameters θ, z, α.

Due to computing constraints we explore a discretized and bounded subset of the θ parameter space. Plausible sections of the θ parameter space were identified by designing a two-stage computer simulation experiment to explore the space. These were labelled Experiment 1 and Experiment 2. Effort expended on precisely estimating θ was minimal since these parameters are not central to the core research question which instead focuses on α. We first describe the experimental design issues considered in defining the discrete parameter space Θ. We then select appropriate prior distributions, a major modelling decision not adressed in the description of EXTENSIONS I AND II. Next, the marginal posterior distribution of each parameter is derived, and a suitable MCMC sampler is chosen. If this choice is a Metropolis-Hastings sampler then the form of the proposal distribution is discussed, and the proposal ratio is computed for the parameter concerned.

Both a frequentist and a Bayesian approach to inference may be taken, depending on the interpretations of results required. The Bayesian approach, however, offers a different and useful approach to interpretation of results, such as deriving the complete marginal distribution of parameters rather than just point and interval estimates. Basing either approach on a hierarchical model permits easy extension of the model on the conceptual plane. Due to the high degree of dependence amongst random variables in the extended models, frequentist approaches to inference are restricted to simple problems, as noted for the basic

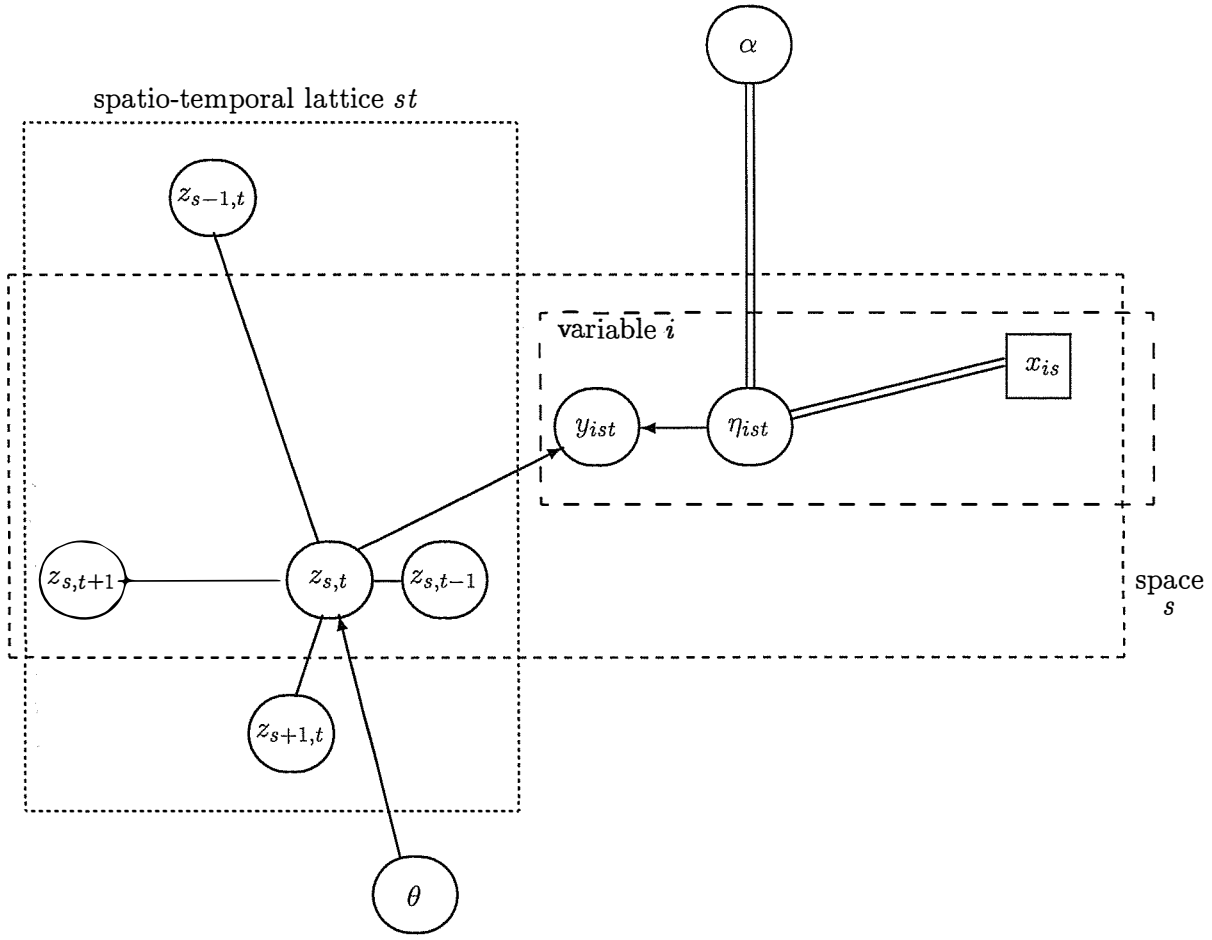


Figure 7.1: Relationship between variables in extension I to hierarchical model for *Dingo* case study, with random θ.

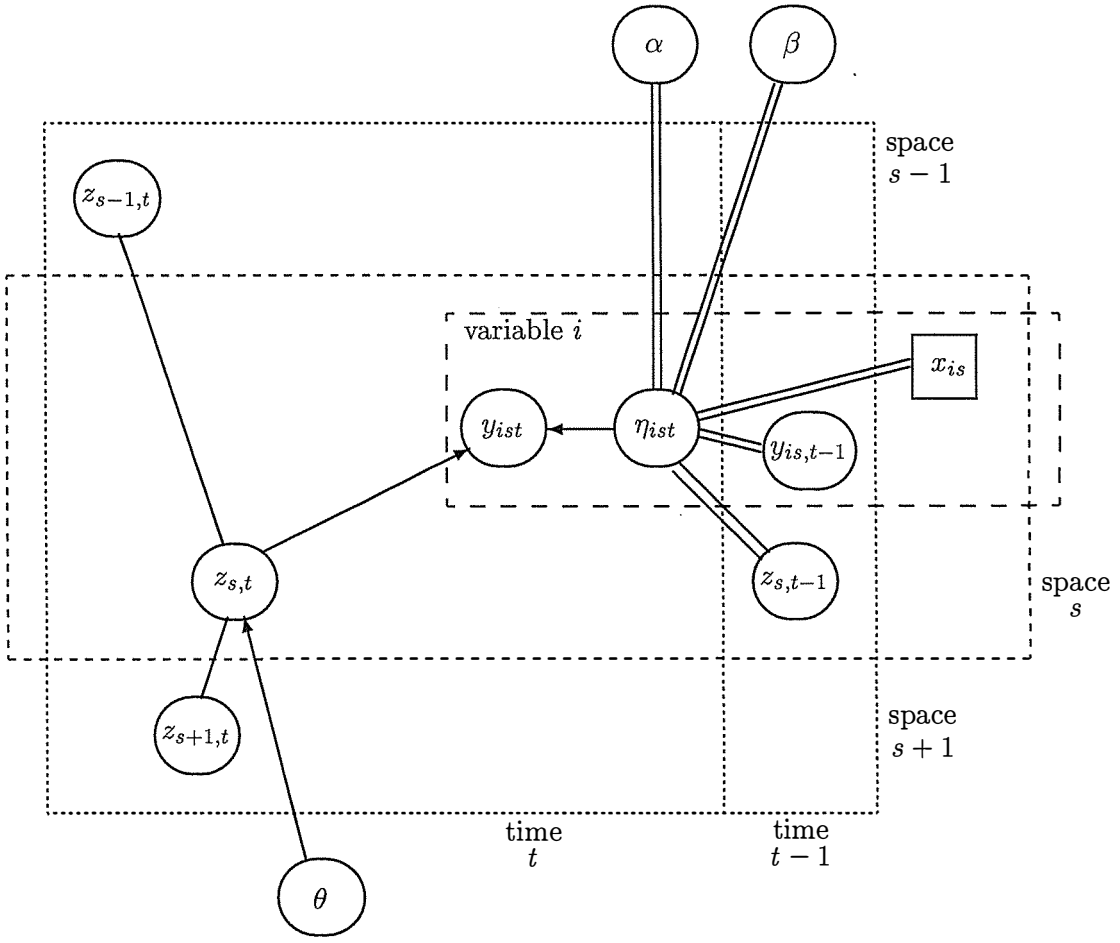


Figure 7.2: Relationship between variables in extension II of hierarchical model for *Dingo* case study, with random θ.

Bayesian model in Chapter 5. We therefore consider application of methods of simulation to inference in the Bayesian context. It is not feasible to obtain independent samples of sets of the random variables from the model, so we instead consider a method of obtaining dependent samples: the Markov Chain Monte Carlo (MCMC) computational algorithm, which is relatively simple to apply to this model. The important steps in implementing an MCMC algorithm for conducting inference on this model were given in more detail in the literature review in Sections 4.4 and 5.4.

In Sections 7.5–7.7, we apply these two extensions to the Bayesian models to the Dingo Case Study as defined in Chapters 2 and further developed in Chapters 3 and 5. Inference proceeds in two stages since computations leading to model inference in EXTENSIONS I AND II both require a small discrete space Θ for spatio-temporal parameter θ. In the first stage a fairly broad resolution space is considered for Θ, and leads to identification of the modal areas of the posterior distribution of θ. A finer resolution Θ space is considered at the second stage.

Finally results from these extended models are discussed in Section 7.8.

7.2 EXTENSION I: extending model for underlying spatio-temporal dependence

This section presents a condensed outline of the distributional assumptions for the first extended model, whose framework is based on that of the original model defined in Chapter 5. Recall that the basic three-tier model conditioned analysis on hyperprior parameters θ, which describe the underlying spatial and temporal dependence. Essentially this is equivalent to placing a prior on θ with point mass at the selected value. In EXTENSION I we instead assume θ is random, and assign a statistical distribution to θ, the hyper-prior $p(\theta|\phi)$.

7.2.1 Data model for y

First we give a specific version of EXTENSION I, suitable for a wide range of applications such as the *dingo* case study. Then we show its relationship to a more general form suitable for modelling data y arising from underlying spatio-temporal dependence z with design matrix X and explanatory variables α. In the general model, the link η between the explanatory variables α and the responses y is made explicit; and the underlying spatial model $p(z|\cdot)$ is no longer restricted to the autologistic distribution, but may be any Markov random field model. For example Weir & Pettitt (1999) use an underlying Gaussian Markov random field, which they term a hidden Gaussian or multivarite probit model.

Dingo case study

In the *dingo* case study, the data model describes the successes/failures y using a Bernoulli distribution, conditioning on the underlying presence/absence process z, the expected probability of success q, and the design factor τ. So the data model is

$$\begin{aligned} \Pr\{y_{ist}|q, z_{ist}=1\} &= q_{\tau_{ist}}^{y_{ist}}(1-q_{\tau_{ist}})^{1-y_{ist}} \\ \Pr\{y_{ist}|q, z_{ist}=0\} &= \text{Ind}[y_{ist}=0]. \end{aligned} \tag{7.1}$$

where the spatial and temporal unit indices range over values $i = 1, \ldots, I$; $s = 1, \ldots, S$; $t = 1, \ldots, T$. Note here that the second line in equation (7.1) embodies a deterministic feature

of the relationship between y and z. Given absence at a site, then it is not possible to observe any successes at the site, so $z_{st} = 0$ implies $y_{ist} = 0$, $\forall i$. Given a success at a site, then there must have been presence at the site, so $y_{ist} \neq 0$ for any i implies $z_{st} = 1$. Given no successes at a site, we encounter the ambiguous zeroes problem however, since it is then unknown whether this corresponds to an absence or presence. The aim of inference is therefore to model $\widetilde{z}_{st}$ in these situations, where $y_{ist} = 0\,\forall i$.

In the frequentist or Bayesian frameworks we may assume that observations y_{ist} were conditionally independent given underlying values of the process z and known covariates x. Thus a joint sampling distribution of all data y is obtained via

$$\begin{aligned} p(y|z,q) &= \prod_{ist} p(y_{ist}|z_{ist},q) \\ p(y|z,q) &= \prod_{ist:z_{ist}=1} q_{\tau_{ist}}^{y_{ist}}(1-q_{\tau_{ist}})^{1-y_{ist}} \cdot \prod_{ist:z_{ist}=0} \mathrm{Ind}[y_{ist}=0]. \end{aligned} \tag{7.2}$$

Note that the last product should always evaluate to unity and therefore cancel for a legal pair of y and z which satisfy the deterministic relationships described in the paragraph above.

Intermediate step towards generalization

The quantity q is the expected mean of y conditional on $z = 1$ and τ:

$$\begin{aligned} q_k &= \mathrm{E}[y_{ist}|z_{ist}=1,\tau_{ist}=k]. \\ 0 &= \mathrm{E}[y_{ist}|z_{ist}=0]. \end{aligned}$$

As for many instances dealing with a probability (Cox 1970) q is best examined on a transformed scale, such as the logit. Set $\tau_{ist} = k$ (as in Chapters 3 and 5) whenever treatment level k is located at unit ist. Then

$$\begin{aligned} \mathrm{logit}(q_k) &= \alpha_k. \\ \mathrm{logit}(q) &= X\alpha. \end{aligned}$$

Hence $\tau_{ist} = k$ if and only if the istth row of X is all 0s except in the kth position where it is 1,*i.e.* $X_{ist,} = [0\ldots010\ldots0]$ where the 1 is in position k.

Generalized Linear Model for y

The above parameterization, although efficient for computations, obscures the structure of the model. This structure becomes more obvious on generalization to a wider range of models for y given z, in the same way that Generalized Linear Models (GLMs) of McCullagh & Nelder (1993) are a generalization of logistic regression. More generally, for any $p(y|z)$ and $p(z|\theta)$, the conditional mean of y given z can be denoted by

$$\mu = \mathrm{E}[y|x,\alpha,z] \tag{7.3}$$

where μ is a $n \times 1$ vector. A linear predictor is formed from the covariates

$$\eta = X\alpha \tag{7.4}$$

where X is the $n \times K$ design matrix and α is a $K \times 1$ vector of explanatory variables. The explanatory variables may be covariates instead of factors in which case the columns of X

each correspond to a different covariate and α is then a column of regression coefficients. Then link function g may be defined to describe the relationship between the responses y and the linear predictor η

$$g(\mu) = \eta; \qquad \mu = g^{-1}(\eta) \tag{7.5}$$

In Chapter 5, we have $\mu \equiv q$, and a logit link $g \equiv \text{logit}$ with inverse $g^{-1}(\eta) \equiv \frac{e^{\eta}}{1+e^{\eta}}$.

Hence using this general notation, the Bernoulli sampling distribution for each data point y_i with generalized index i becomes

$$p(y_i|z_i,\alpha) = \begin{cases} \left(\frac{e^{\eta_i}}{1+e^{\eta_i}}\right)^{y_i} \left(\frac{1}{1+e^{\eta_i}}\right)^{1-y_i}, & z_i = 1; \\ 1 - y_i & z_i = 0; \end{cases} \tag{7.6}$$

where $\eta_i = (X\alpha)_i = \sum_k x_{ik}\alpha_k$ is the ith component of the linear predictor. Although this expression is useful for the general formulation, the simplified expression 7.1 is more useful for computations since it takes advantage of the structure of the design matrix X.

The joint sampling distribution of all the data is then

$$p(y|z,\alpha) = \prod_{i:z_i=1} \frac{e^{\eta_i y_i}}{1+e^{\eta_i}} \cdot \prod_{i:z_i=0} \text{Ind}[y_i = 0] \tag{7.7}$$

The Bernoulli-Autologistic framework can therefore be extended so that y may have any distribution in the exponential family (*e.g.* Binomial, Poisson, Gaussian) and the underlying spatio-temporal process z may have an appropriate Markov Random Field distribution $p(z|\theta)$ (*e.g.* auto-binomial, auto-Poisson, auto-Gaussian).

7.2.2 Underlying spatio-temporal dependence z

Within a general modelling framework, literature reviewed in Chapter 4 suggests that the underlying spatio-temporal dependence process z can be modelled by a Markov Random Field where:

$$\begin{aligned} h(z|\theta) &= \exp\{\theta^{\mathsf{T}} V(z)\} \\ p(z|\theta) &= \frac{h(z|\theta)}{c(\theta)} \\ c(\theta) &= \textstyle\sum_{z\in\Omega} h(z|\theta) \end{aligned} \tag{7.8}$$

In the *dingo* case study, underlying presence/absence was modelled using a 3-parameter autologistic distribution, with parameters determining prevalence and spatial and temporal dependence fixed within a realm suggested by Experts.

$$\theta = (\theta_0, \theta_1, \theta_2)$$

$$V(z) = \begin{bmatrix} \sum z_{st} \\ \sum z_{st} z_{s-1,t} \\ \sum z_{st} z_{s,t-1} \end{bmatrix}$$

In the general modelling framework, the form of θ and V in (7.9) may be replaced by any of those discussed in Chapter 4.

7.2.3 Joint distribution

We can assume that the effects α and associated explanatory variables X are independent of the presence/absence process z and associated parameters θ and ϕ. This is somewhat analogous to assumptions in linear models where error distributions are considered independent of parameters contributing to linear predictors (McCullagh & Nelder 1993); in repeated measures temporal dependence between measurements on the same unit is considered independent of other parameters in the model (Diggle, Liang & Zeger 1996).

As part of the Bayesian inferential process, we require a prior distribution $p(\alpha)$ for the effects α. If there is a one-to-one mapping between α and μ then it is equivalent to apply the prior on the scale of the covariates (α) or on the scale of the data (μ). Thus the full joint distribution of all parameters, conditioning only on ϕ is

$$p(y, x, \alpha, z, \theta|\phi) = p(y|x, \alpha, z)p(\alpha)p(z|\theta)p(\theta|\phi). \tag{7.9}$$

A key property of Markov Random Fields such as the autologistic distribution ensures that the joint distribution $p(z|\theta)$ implicit in equation (7.9) can be expressed as the product of local relationships $p(z_i|\theta)$. This property is known as the equivalence theorem for Markov Random fields and Gibbs Random fields (page 74.)

In the *dingo* case study, the basic model permitted us to make probabilistic statements about the relative probabilities of any dingoes visiting particular chemicals, provided dingoes were present in the vicinity. Thus treatments (chemicals) could be ranked in order of how successful they were. This basic model, however, stopped short of permitting inference on the absolute probabilities of success (visits) given presence. It was not possible to determine just how successful each treatment (chemical) was, which would be imperative for any cost-benefit analysis deciding whether to use a particular treatment (chemical).

7.2.4 Prior for spatio-temporal dependence θ

It is at this point that we encounter that aspect of EXTENSION I which distinguishes it from the basic model: the addition of another layer $p(\theta|\phi)$ describing the distribution of underlying spatial dependence parameters θ. The extra layer was found (on page 2.3.1) to be necessary because expert opinion could not accurately assign values to θ. This highlights some difficulties with relying on expert opinion: insufficient information; or as in the *dingo* case study, the experts have differing opinions! Some animal behavourial scientists believed that the dingoes would travel along the road since it provided an easy path through their home range; other scientists believed that the road would have little impact on their movements. This is a situation where modelling the extra hyper-prior parameters could therefore contribute to scientific knowledge. This modelling framework makes it possible to evaluate which scientific theory on underlying dingo movement patterns is best supported by the data observed.

Thus in EXTENSION I θ is no longer a fixed parameter for the prior $p(z|\theta)$, but is itself considered random, and can be used to cover the gamut of expert opinion, with

$$\theta \sim p(\theta|\phi) \tag{7.10}$$

where ϕ is a vector of parameters driving spatio-temporal dependence. As explained in more detail later in Section 7.2.10, the trouble with incorporating equation (7.10) is that

computationally it is no longer possible to avoid evaluating ratios of Normalization Constants

$$\lambda(A,B) = \frac{c(\theta_A)}{c(\theta_B)} = \frac{\sum_{z\in\Omega} h(z|\theta_A)}{\sum_{z\in\Omega} h(z|\theta_B)}. \tag{7.11}$$

This problem of evaluating Normalization Constant ratios was addressed in Chapter 6. It facilitates the computations, based on MCMC, required to estimate the extra hyper-prior parameters for this extension to the model.

7.2.5 Hyperparameters ϕ

There are various choices for $p(\theta|\phi)$ and ϕ, which depend to a large extent on the application. A common choice would be a uniform $p(\theta|\phi)$ which reflects the situation where no value of θ is favoured over any other. In this case ϕ represents the interval of the uniform distribution, namely the range of plausible spatio-temporal dependence parameters. Other choices include a Gaussian or triangular distribution centred around plausible values of θ.

A major decision is whether spatial or temporal competition should be permitted: should the presence at a particular space-time site inhibit presence at neighbouring sites? If this is the case, then negative values of θ_1, θ_2 are possibilities which need to be incorporated into the hyperprior. This question was only addressed briefly in Chapter 4. Furthermore is the lattice a closed system, where the absolute number of presences/absences over the entire space-time lattice is fixed, as for gas particles in a closed chamber, or molecules in a crystal? This affects θ_0, and may simplify simulation of the model (as per Section 4.4.5) and therefore computations leading to inference.

In the *dingo* case study, the choice of ϕ is driven by several practical constraints. Firstly the number of observed y, z was not high so does not warrant complex hyperprior models. Thus the form of $p(\phi)$ should be simple. Secondly conflicting information from Experts on dingo spatial movements suggested that ϕ should cover a variety of possible spatio-temporal patterns, with spatial dependence ranging from non-existent ($\theta_1 = 0$) to high and positive ($\theta_1 > 0$) and similarly for temporal dependence. It was found to be simpler in practice to consider the sum of spatio-temporal dependence, $(\theta_1 + \theta_2)$, and the comparison between these, $(\theta_1 - \theta_2)$. Thirdly Experts did agree *a priori* that in their natural habitat dingoes within the same pack tend to range widely (Corbett 1995). Furthermore, dingoes in different packs do not tend to interact (Corbett 1995). Thus the proportion of space-time sites visited by dingoes in any pack (or lone animals) tends to be low. All three constraints on parameters can be used to fix the most extreme values of each component of θ, and then a simple choice for $p(\theta|\phi)$ is a discrete uniform distribution over this restricted θ-space. Extreme values of components $\theta_0, \theta_1, \theta_2$ satisfying these constraints were discussed in detail in Chapter 5.

At this stage we can opt to consider each element of θ independently or as a vector. By selecting the vector representation we ensure that the tradeoffs between prevalence θ_0 and dependence (θ_1, θ_2) are not ignored. In this phase of the *dingo* case study, I selected the vector option which essentially models the full spatio-temporal dependence simultaneously.

7.2.6 Prior for coefficients of covariates, α

Setting a prior for regression coefficients can be achieved either on the scale of the responses or on the scale of the linear predictor. For example a uniform prior for q_k implies that all probabilities between 0 and 1 are equi-likely in the absence of specific prior information.

In contrast a uniform prior for α_k signifies that probabilities in the middle range (*e.g.* 0.15 to 0.85) are approximately equi-likely, but the likelihood of more extreme probabilities decreases as a function of order, rather than magnitude, due to the logit transformation. Furthermore since α is not restricted to the unit interval, it can be modelled by a distribution with support on $\Re$.

The choice of prior distribution and scale of application depends on the application, and the existence of prior substantive knowledge. When there is no prior information the coefficients α can be allocated a uniform prior; when some prior information exists on ranges of possible values these can be built into a normal prior centred on the most probable value (*a priori*), with variance indicating the range of plausible values.

A major decision is whether to model the $\{\alpha_k\}$ as fixed or random effects. We would decide in favour of the latter option if the α components can be considered exchangeable so that $\alpha_k \sim \text{IID N}(a, \Sigma)$, providing a suitable prior. Alternatively if the α components are considered to be fixed effects then this can either be represented by $\Sigma = \infty$ in the random effects distribution.

In the *dingo case study*, prior information on the efficacy of different treatments could have been obtained from results of the pen trials. This could have been a viable choice if it was thought that behaviour of wild dingoes in the field study would be comparable to the behaviour of captured dingoes in the pen trial studies. Recall that one of the underlying research questions to be addressed by the field study was whether these two behaviours were similar or not, so this prior information would make a large contribution to results, which would then reflect the information gained from the pen trials. Note that the extent of the contribution would be fairly small, since the the data have a large impact on the posterior distribution of α.

Thus, in deliberate ignorance of prior information, the uniform prior on either the linear predictor scale or on the observed scale would be appropriate. Another factor to consider in this case study is that it was noted in the pen trials that there was up to an order of magnitude difference between the treatment effects (chemical attractiveness). A uniform prior on the scale of the linear predictor would be better able to capture this range of scales in the effects compared to the same prior on the scale of the observations.

7.2.7 Posterior distributions

Posterior distributions cannot be derived without assuming some form for the prior distributions, in addition to the data models. In the following sections, we take the prior forms suitable for the *dingo* case study. These forms are sufficiently general since they are based on an assumption that *a priori* there is no information to justify favouring some parameter values over any others.

Some economy of notation is achieved by retaining the notation of Chapter 5, hence equations are presented in this specific form first before presenting the more general form of expressions (equation (7.3))–(equation (7.5)).

7.2.8 Marginal Posterior distribution for α

As discussed in Chapter 5 the posterior distribution of probability of success q_k only involves those sites with treatment k:

$$p(q_k|\ldots) \propto \prod_{is:\tau_{is}=k} \prod_t p(y_{ist}|q_k, z_{st}) p(q_k). \tag{7.12}$$

Hence the posterior ratio, that is the ratio of the posterior distribution evaluated at two values of q_k, denoted q'_k and q_k, is

$$\begin{aligned}\frac{p(q'_k|\ldots)}{p(q_k|\ldots)} &= \frac{p(q'_k)}{p(q_k)} \prod_{is:\tau_{is}=k} \prod_t \frac{p(y_{ist}|q'_k, z)}{p(y_{ist}|q_k, z)} \\ &= \frac{p(q'_k)}{p(q_k)} \left(\frac{q'_k}{q_k}\right)^{S(k11)} \left(\frac{1-q'_k}{1-q_k}\right)^{S(k10)}\end{aligned} \tag{7.13}$$

where

$$S(kzy) = \sum_{ist} \text{Ind}[\tau_{is} = k,\ z_{st} = z,\ y_{ist} = y] \tag{7.14}$$

the sum over all sites with the kth treatment, presence/absence z, and success/failure y.

7.2.9 Marginal Posterior distribution for z_{st}

The joint posterior for imputing missing values of z denoted by $z \in \widetilde{\Omega}$ depends on the likelihood and on the autologistic model for z:

$$p(z|\ldots) \propto p(y|q, z)p(z|\theta). \tag{7.15}$$

The posterior for individual $z_{st} \in \widetilde{\Omega}$ is

$$p(z_{st}|\ldots) \propto p(z_{st}|z_{-st}, \theta) \prod_i p(y_{ist}|q_{\tau_{is}}, z_{st}) \tag{7.16}$$

where

$$p(z_{st}|z_{-st}, \theta) = \frac{\exp\{z_{st}\theta^{\mathsf{T}} V_{st}(z)\}}{c_{st}(\theta)}$$

$$V_{st}(z) = \begin{bmatrix} 1 \\ z_{s-1,t} + z_{s+1,t} \\ z_{s,t-1} + z_{s,t+1} \end{bmatrix} \tag{7.17}$$

$$\begin{aligned} c_{st}(\theta) &= \textstyle\sum_{z_{st}=0,1} \exp\{z_{st}\theta^{\mathsf{T}} V_{st}(z)\} \\ &= 1 + \exp\{\theta^{\mathsf{T}} V_{st}(z)\} \end{aligned}$$

or odds of a success compared to a failure at site st are given by

$$\frac{p(z_{st} = 1|z_{-st}, \theta)}{p(z_{st} = 0|z_{-st}, \theta)} = \exp\{\theta^{\mathsf{T}} V_{st}(z)\} \tag{7.18}$$

Thus in the posterior distribution for z_{st}, the odds ratio for comparing $z'_{st} = 1$ to $z'_{st} = 0$ can therefore be obtained exactly (which later allows use of a simple Gibbs update):

$$\frac{p(z'_{st} = 1|\ldots)}{p(z'_{st} = 0|\ldots)} = \prod_i p(y_{ist}|q_{\tau_{ist}}, z_{st} = 1) \exp\{\theta^{\mathsf{T}} V_{st}(z)\} \tag{7.19}$$

7.2.10 Marginal Posterior distribution for θ

The posterior distribution for Presence dependence θ changes with the extensions to the hierarchical model due to the additional level incorporated to describe its distribution $p(\theta)$. It depends only on the underlying presence model and the prior distribution for θ:

$$p(\theta|\ldots) \propto p(z|\theta)p(\theta|\phi). \tag{7.20}$$

The ratio of posteriors, comparing θ' to θ, involves the ratio of normalization constants, and is given by

$$\frac{p(\theta'|\ldots)}{p(\theta|\ldots)} = \exp\{(\theta'-\theta)^{\mathsf{T}}V(z)\}\frac{c(\theta)}{c(\theta')}\frac{p(\theta')}{p(\theta)}\frac{r(\theta'|\theta)}{r(\theta|\theta')} \tag{7.21}$$

The real challenge here is the unknown normalization constant ratio $\frac{c(\theta)}{c(\theta')}$. Recall the expression for the NC is:

$$c(\theta) = \int_{z\in\Omega} \theta^{\mathsf{T}}V(z) \tag{7.22}$$

Since Ω is ordinarily very large (*e.g.* in the dingo example $\Omega = 2^{700}$) this sum is quite infeasible to compute. This complex computational though simply posed problem has been discussed in detail in Chapter 6. Here we simply apply the results obtained from the other chapter's in-depth focus on the computational problem. Briefly we mention the failure of enumeratirve and simple Monte Carlo methods and the limitations of analytic approximations in attempting to solve the problem. We found the Reverse Logistic Regression (Geyer 1994, Geyer 1996) to work adequately. A much quicker estimating equation method, however, is based on integrating the mean canonical statistical and has comparable precision to the Reverse Logistic Regression method. Both of the last-mentioned methods are used to ensure accuracy.

New θ^* values can be generated from a discrete uniform proposal distribution centred on the current value of θ in 3-dimensional space, and extending up to two "jumps" in any direction.

Starting values for θ were chosen from the discrete uniform distribution over the range of $\{\theta_m\}$ selected for the experiment.

7.2.11 Computational issues: MCMC design

The q_k parameters

A truncated uniform proposal distribution on the uniform interval, based at the current value q_k, with defined half-width h was chosen for the new proposed value q_k' for the posterior distribution in Section 7.2.8. Other choices for a proposal distribution are a rotated uniform distribution or a truncated normal distribution.

Values between 0 and 1.0 for the starting value $q^{(0)}$ should be assessed for their impact on convergence. A value of 0.3 should especially be investigated since it appeared to be associated with quick convergence in the basic model.

The unknown z_{st} parameters

A Gibbs update is suitable for updating z_{st} since its posterior distribution can be written down precisely, and is easily sampled from. Thus a proposal distribution is not required.

Starting values for z of either 0 and 1 should also be examined. A Bernoulli distribution with probability of a presence 0.2 was selected in the basic model for its association with quick convergence. (Recall that the results of Section 5.5.5 from the pilot study of the *dingo* example suggested that a random method of assigning starting values to the components of z worked best of the four options considered.) The value of 0.2 represents an upper bound on the prior median overall prevalence on the lattice.

The θ parameters

We select a Metropolis-Hastings sampler for updating θ since it is simpler to write down ratios of the posterior distributions, which involve ratios of Normalization Constants, rather than the complete posterior distribution, which involves the raw Normalization Constant. In practice, a finite but large number of θ values will be compared. Therefore a Metropolis-Hastings update is more efficient than a Gibbs sampling approach.

New θ' values can be generated from a rotated discrete uniform proposal distribution centred on the current value of θ in 3-dimensional space, and extending up to h "jumps" of size δ_l in any direction l. (For simplicity we assume that the discretized sample space Θ has points equally spaced in each dimension.) This version of a rotated uniform is merely the discrete three-dimensional version of equations (4.4.2)–(4.114). Recall from Section 4.4.2 a rotated uniform is just a uniform distribution wrapped around the unit interval, so that the neighbour of 1 is 0 and vice versa.

Thus the component-wise proposal distribution is:

$$\begin{aligned} \theta_l' \,|\, \theta_l &\sim \text{Discrete Uniform for } \theta_l' \in \Theta_l \text{ wrapped onto } [\Theta_l^{\mathsf{L}}, \Theta_l^{\mathsf{U}}]. \\ \Theta_l &= [\theta_l - h_l\delta_l,\ \theta_l - (h_l+1)\delta_l,\ \ldots,\ \theta_l,\ \ldots, \\ &\qquad \theta + (h_l-1)\delta_l,\ \theta + h_l\delta_l] \end{aligned} \tag{7.23}$$

where the wrapping implies that the neighbour of the edges of the proscribed interval are given by the opposite edge, *i.e.* the left neighbour of Θ_l^{L} is Θ_l^{U} and vice versa. Now the proposal distribution for vector θ is easily obtained (when assuming that each component is proposed independently of the others):

$$r(\theta'|\theta) = \prod_l r(\theta_l'|\theta_l) \tag{7.24}$$

The proposal ratio is therefore

$$\frac{r(\theta'|\theta)}{r(\theta|\theta')} = 1, \qquad \text{for } \theta' \in \prod_l [\Theta_l^{\mathsf{L}}, \Theta_l^{\mathsf{U}}]. \tag{7.25}$$

Starting values for θ in a MCMC computational framework can be chosen from the discrete uniform distribution of equation (7.23) over the range of Θ selected for the experiment.

7.3 EXTENSION II: Modelling time-dependent success rather than time-dependent presence

The second extension of the basic model in Chapter 5 allows time dependence to appear in the data model for success (given presence) rather than in the sub-model for underlying

presence/absence. It is an alternative to EXTENSION I of the basic model presented above in Section 7.2.

7.3.1 Data model for y

First a general form of the data model is given for each component y_{st} of y. In order to achieve this the general index i used in Section 7.2 is split so that the indices distinguish between spatial position s and temporal position t. (To directly apply the following results to the *Dingo* case study, just replace s by the tuple is.) Secondly these component-wise data models are combined to construct the complete data model for all components of y. Finally, some properties of the probability transition matrix defined by this model are presented. These are found to be useful later.

General formulation for y_{st}

We consider that in the expression for the probability of success given presence, time dependence could enter as an auto-regressive temporal effect, β.

$$\text{logit}\,(\mu_{st}) = \eta_{st} = X_{st}\alpha + y_{s,t-1}z_{s,t-1}\beta \tag{7.26}$$

where X_{st} denotes the stth row of design matrix X. Note that with this parameterization β only enters if there was success at the previous time point. In the *dingo* case study $X_{ist,j} = 0$ for all $j \neq k$ where k designates which treatment was assigned to unit ist. This is referred to as a "transition model" (Diggle et al. 1996, Ware 1985).

In the notation of the *dingo* case study, this is equivalent to

$$\text{logit}\,(q_{\tau_{ist}}) = \begin{cases} \alpha_{\tau_{ist}} + \beta y_{is,t-1}, & z_{s,t-1} = 1 \\ \alpha_{\tau_{ist}}, & z_{s,t-1} = 0 \end{cases} \tag{7.27}$$

Let us consider how the autoregressive effect parameter β can be interpreted. In this model there is a 'carryover' effect, meaning that if the previous success affects the probability of success far more than the effect of the treatments α_k, then this would be evidenced by $|\beta| >> |\alpha_k|$. A positive value of β would imply that the probability $p(y_{ist} \,|\, y_{is,t-1}, z_{st} = 1, z_{s,t-1} = 1)$ would be large and close to 1, whereas a negative value of β would imply that this probability would be close to 0.

7.3.2 Joint data model for y

The joint distribution of y components at the same site can be written in terms of conditional distributions with respect to time:

$$\begin{aligned} p(\{y_{st},\, t = 1,\ldots,T\} \,|\, \ldots) \;=\; & p(y_{s1} \,|\, \ldots)\, p(y_{s2} \,|\, y_{s1}; \ldots)\, p(y_{s3} \,|\, y_{s1}, y_{s2}; \ldots) \;\times \\ & \ldots \times p(y_{sT} \,|\, y_{s1}, y_{s2}, \ldots, y_{s,T-1}; \ldots). \end{aligned}$$

We make an assumption that the likelihood at time t for a particular site s follows a first-order Markov structure and is conditional on the previous time period only. So the joint likelihood at site s simplifies to

$$\begin{aligned} p(\{y_{st},\, t = 1,\ldots,T\} \,|\, \ldots) \;=\; & p(y_{s1} \,|\, \ldots)\, p(y_{s2} \,|\, y_{s1}; \ldots)\, p(y_{s3} \,|\, y_{s2}; \ldots) \;\times \\ & \ldots \times p(y_{sT} \,|\, y_{s,T-1}; \ldots). \end{aligned}$$

The relationship between components y_{st} in the joint likelihood of the data y, given the underlying spatio-temporal process Z and explanatory variates X, also reflects the autoregressive nature of the model:

$$p(y \mid z, \alpha, \beta) = \prod_s \left\{ p(y_{s1} \mid z, \alpha, \beta) \prod_{t \geq 2} p(y_{st} \mid z, y_{s,t-1}, \alpha, \beta) \right\} \tag{7.28}$$

$$p(y \mid z, \alpha, \beta) = \prod_s \left\{ p(y_{s1} \mid z_{s1}, \alpha, \beta) \prod_{t \geq 2} p(y_{st} \mid z_{st}, y_{s,t-1}, \alpha, \beta) \right\} \tag{7.29}$$

This likelihood is essentially an autoregressive logistic model for y based on a linear predictor η

$$p(y_{st} = 1 \mid z, y_{s,t-1}) = \frac{e^{\eta_{st}}}{1 + e^{\eta_{st}}} \tag{7.30}$$

where η is given in (7.26).

Due to the Bernoulli nature of the distribution $p(y_{st} \mid z, y_{s,t-1}, \alpha, \beta)$ it is not possible to separate the auto-regressive error from the residual error after accounting for covariates.

A difficulty with an auto-regressive model in time is that the distribution of the initial time period must be specified. A reasonable option is to assume that all sites have hitherto been unsuccessful (*e.g.* sites are fresh and were not visited in initial time period just before chemicals placed.) Hence in this situation $y_{is0} = 0 \; \forall i, s$. Then the distribution of the first day can be written as

$$p(y_{is1}) = \frac{e^{X_{is1}\alpha}}{1 + e^{X_{is1}\alpha}} \tag{7.31}$$

where the effect of the previous time period is eliminated due to the enforced failures in that time period. However in the *dingo* case study this was not the case since the results from the first day of eight were discarded as they suffered from problems with completeness. So in this type of situation, the distribution of the first day can be used as a baseline using the same expression as above in equation (7.31). The distribution of subsequent days incorporate an effect $\beta y_{is,t-1}$ which can, in this situation, be interpreted as the *additional* effect of the previous day's success compared to the baseline given by the first day.

Properties of the probability transition matrix

In any autoregressive model such as this one, it is of interest to see what the distribution of y will tend to eventually in equilibrium as $t \to \infty$, given that all sites have presence, *i.e.* $z_{st} = 1 \forall s, t$ which we will write as $z = 1$. The probability transition matrix $p(y_{st} \mid y_{s,t-1}, z = 1)$ given that treatment k has been applied at the site s is tabled below.

$$\begin{array}{cc} & y_{st} \\ & \begin{array}{cc} 0 & 1 \end{array} \\ y_{s,t-1} \begin{array}{c} t \\ t-1 \end{array} & \left[\begin{array}{cc} \frac{1}{1+e^{\alpha}_k} & \frac{e^{\alpha}_k}{1+e^{\alpha}_k} \\ \frac{1}{1+e^{\alpha_k+\beta}} & \frac{e^{\alpha_k+\beta}}{1+e^{\alpha_k+\beta}} \end{array}\right] \equiv \left[\begin{array}{cc} 1-a_k & a_k \\ b_k & 1-b_k \end{array}\right] \end{array} \tag{7.32}$$

where each component P is $P_{uv} = p(y_{st} = v \mid z = 1, y_{s,t-1} = u)$, $u, v = 0, 1$ and a_k and b_k are the probabilities of switching in time steps;

$$\begin{aligned} a_k &= \frac{e_k^{\alpha}}{1 + e_k^{\alpha}} = p(y_{st} = 0 \mid z = 1, y_{s,t-1} = 1, \alpha) \\ b_k &= \frac{1}{1 + e^{\alpha_k + \beta}} = p(y_{st} = 1 \mid z = 1, y_{s,t-1} = 0, \alpha, \beta) \end{aligned}$$

The limiting distribution in equilibrium for this probability transition matrix is

$$\pi(y_{st} \mid z = 1) = \begin{bmatrix} \pi(y_{st} = 0 \mid z = 1) \\ \\ \pi(y_{st} = 1 \mid z = 1) \end{bmatrix} = \begin{bmatrix} \dfrac{b_k}{a_k + b_k} \\ \\ \dfrac{a_k}{a_k + b_k} \end{bmatrix} \tag{7.33}$$

Thus the limiting odds of success to failure at all sites (if presence is assumed at all sites) is just

$$\frac{p(y_t = 1 \mid \alpha)}{p(y_t = 0 \mid \alpha)} = \frac{a_k}{b_k}. \tag{7.34}$$

The odds of presence to absence from the equilibrium distribution compared for treatment $k = 1$ versus treatment $k = 2$ are given by

$$\frac{a_1/b_1}{a_2/b_2} = \left(\frac{\frac{e^{\alpha_1}}{1+e^{\alpha_1}} \Big/ \frac{1}{e^{\alpha_1+\beta}}}{\frac{e^{\alpha_2}}{1+e^{\alpha_2}} \Big/ \frac{1}{e^{\alpha_2+\beta}}} \right) \tag{7.35}$$

7.3.3 Underlying presence/absence model Z

In the underlying model for presence (Z) for the basic model, time dependence enters as θ_2 the temporal effect

$$p(z \mid \theta) \equiv p(z \mid \theta_0, \theta_1, \theta_2)$$

In this approach the temporal effect is captured (above) in the term β in the data model, so it is necessary to elimininate the θ_2 parameter from the underlying spatial presence/absence process

$$p(z \mid \theta) \equiv p(z \mid \theta_0, \theta_1)$$

where

$$\begin{aligned} \theta &= (\theta_0, \theta_1) \\ V(z) &= \begin{bmatrix} \sum_{st} z_{st} \\ \sum_{st} z_{st} z_{s-1,t} \end{bmatrix}. \end{aligned} \tag{7.36}$$

7.3.4 Prior for parameters of Presence/absence θ

Compared to the prior selected for θ for EXTENSION I the only difference is that θ now only has two components instead of three. Refer to Section 7.2.4 for discussion on the prior distribution.

7.3.5 Hyperprior parameters ϕ

Again refer to prior selected for ϕ for EXTENSION I in Section 7.2.5.

7.3.6 Prior for chemical attractiveness parameters α_k

The prior for α now involves the auto-regressive parameter β and so differs from EXTENSION I. Before selecting the prior distribution for these parameters, we discuss broad prior requirements.

For presence, in the absence of presence at neighbouring sites then the conditional (local) probability of presence is given by

$$p(z_{st} \mid z_{s-1,t} = 0, z_{s+1,t} = 0, z_{s,t-1} = 0, z_{s,t+1} = 0) = \frac{e^{\theta_0}}{1 + e^{\theta_0}} \tag{7.37}$$

In the dingo case study this probability is likely to be in the proximity of $[0.10, 0.15]$, so $\theta_0 = \ln\left(\frac{p}{1-p}\right)$ is likely to be within the range $[-2.2, -1.7]$.

For successes conditional on presence and given that there was a failure the previous day

$$p(y_{ist} = 1 \mid y_{is,t-1} = 0, z_{st} = 1, z_{s,t-1} = 0) = \frac{e^{\alpha_k}}{1 + e^{\alpha_k}} \tag{7.38}$$

where $k = \tau_{is}$. For $\alpha_k = 0$, $p = \frac{1}{2}$. Obviously α_k should be adjusted to reflect this conditional probability for the particular application being considered.

Some possible choices for a prior distribution for α_k are Normal or Logistic. With a normal prior, the prior distribution for α_k is:

$$p(\alpha_k^* \mid \sigma_\alpha, \mu_\alpha) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{(\alpha_k^* - \mu_{\alpha_k})^2}{2\sigma_{\alpha_k}^2}\right\}$$

With a Normal prior, the ratio of priors is straightforward to compute:

$$\frac{p(\alpha_k^*)}{p(\alpha_k)} = \exp\left\{\frac{\left(\alpha_k^{*2} - \alpha_k^2\right) - \mu_{\alpha_k}(\alpha_k^* - \alpha_k)}{2\sigma_{\alpha_k}^2}\right\} \tag{7.39}$$

Another option is to let α_k have a logistic prior distribution, in which case the ratio of priors is:

$$\begin{aligned} \frac{p(\alpha_k^*)}{p(\alpha_k)} &= e^{(\alpha_k^* - \mu_\alpha - \alpha_k + \mu_\alpha)} \frac{\left(1+e^{(\alpha_k - \mu_\alpha)}\right)^2}{\left(1+e^{(\alpha_k^* - \mu_\alpha)}\right)^2} \\ &= e^{(\alpha_k^* - \alpha_k)} \frac{\left(1+e^{(\alpha_k - \mu_\alpha)}\right)^2}{\left(1+e^{(\alpha_k^* - \mu_\alpha)}\right)^2} \end{aligned} \tag{7.40}$$

7.3.7 Prior for time dependence parameter β

We would wish the prior distribution for β to reflect our prior expectation that the time effect could be either positive or negative, but that it is most likely not very large in magnitude. A standardized Normal distribution truncated to the interval $[\beta_L, \beta_U]$ appears to provide enough flexibility without being too restrictive. Thus the distribution of β^* is

$$p(\beta^* \mid \sigma_\beta, \mu_\beta) = \frac{1}{\Phi\left(\frac{\beta_U - \mu_\beta}{\sigma_\beta}\right) - \Phi\left(\frac{\beta_L - \mu_\beta}{\sigma_\beta}\right)} \frac{1}{\sqrt{2\pi}\sigma_\beta} \exp\left\{\frac{(\beta^* - \mu_\beta)^2}{2\sigma_\beta^2}\right\}$$

Similarly to equation (7.39), since the normalizing constant cancels, the ratio of prior distributions in this case is therefore

$$\frac{p(\beta^*)}{p(\beta)} = \exp\left\{\frac{\left(\beta^{*2} - \beta^2\right) - \mu_\beta\left(\beta^* - \beta\right)}{2\sigma_\beta^2}\right\}$$

7.3.8 Marginal Posterior distributions

Marginal Posterior for q_k

The posterior distribution for the treatment effects α_k is given by

$$p(\alpha_k \mid \ldots) \propto \underbrace{p(\alpha_k)}_{\text{TERM 1}} \times \underbrace{\prod_{is} p(y_{is1} \mid q_{is1}, z_{st})}_{\text{TERM 2}} \times \underbrace{\prod_{is}\prod_{t=2}^{T} p(y_{ist} \mid y_{is,t-1}, z_{st}, z_{s,t-1}, q_{ist})}_{\text{TERM 3}} \tag{7.41}$$

We will consider each of the terms separately. TERM 1 is the prior distribution, discussed in the previous section. When working with binary variables, there are a number of equivalent ways of writing expressions involving these variables, *e.g.* logical expressions and products. When taking products over expressions involving the binary variables care needs to be taken to ensure that the expression evaluates to 1 in some situations. Some expressions made use of in this chapter are as follows.

	evaluated for	
expression	$z = 0$	$z = 1$
$[zq + (1-z)]^y$	1	q^y
$(1 - zq)^{1-y}$	1	$(1-q)^{1-y}$
$\left(\frac{ze^\alpha}{1+e^\alpha} + (1-z)\right)^y$	1	$\left(\frac{e^\alpha}{1+e^\alpha}\right)^y$
$\left(\frac{1+(1-z)e^\alpha}{1+e^\alpha}\right)^{1-y}$	1	$\left(\frac{1}{1+e^\alpha}\right)^{1-y}$

The portions of TERM 2 that involve α_k are

$$\begin{aligned}\text{TERM 2} \propto\ & \textstyle\prod_{is:\tau_{is}=k} \left(z_{s1}q_{is1} + (1 - z_{s1})\right)^{y_{is1}} \left(1 - z_{s1}q_{is1}\right)^{1-y_{is1}} \\ =\ & \underbrace{\textstyle\prod_{is:\tau_{is}=k} \left(\frac{z_{s1}e^{\alpha_k}}{1+e^{\alpha_k}} + (1-z_{s1})\right)^{y_{is1}}}_{\text{TERM 2a}} \times \underbrace{\left(\frac{1 + (1-z_{s1})e^{\alpha_k}}{1+e^{\alpha_k}}\right)^{1-y_{is1}}}_{\text{TERM 2b}}\end{aligned}$$

where

$$\text{TERM 2a} = \frac{(e^{\alpha_k})^{\sum_{is} I[\tau_{is}=k, y_{is1}=1]}}{(1+e^{\alpha_k})^{\sum_{is} I[\tau_{is}=k, z_{s1}=1]}}.$$

Note that $y_{is1} = 1$ only if $z_{s1} = 1$, which simplifies the numerator above. Considering the two cases $z_{s1} = 0$ and $z_{s1} = 1$ separately

$$\text{TERM 2b} = \left(\frac{1+e^{\alpha_k}}{1+e^{\alpha_k}}\right)^{\sum_{is} I[\tau_{is}=k, y_{is1}=0, z_{s1}=0]} \times \left(\frac{1}{1+e^{\alpha_k}}\right)^{\sum_{is} I[\tau_{is}=k, y_{is1}=0, z_{s1}=1]}.$$

Therefore combining these we obtain

$$\text{TERM 2} = \frac{\left(e^{\alpha_k}\right)^{\Sigma_1}}{\left(1+e^{\alpha_k}\right)^{\Sigma_2}}$$

where

$$\begin{aligned}
\Sigma_1 &= \textstyle\sum_{is} I\left[\tau_{is} = k, y_{is1} = 1, z_{s1} = 1\right] \\
\Sigma_2 &= \textstyle\sum_{is} I\left[\tau_{is} = k, y_{is1} = 1\right] + \sum_{is} I\left[\tau_{is} = k, y_{is1} = 0, z_{s1} = 1\right] \\
&= \textstyle\sum_{is} I\left[\tau_{is} = k, z_{s1} = 1\right].
\end{aligned}$$

Note that in the above, that if $y_{ist} = 1$ then $z_{st} = 1$ necessarily, however it is sometimes useful to continue writing both statements for comparison with other cases. Now the third term can also be written in a similar way:

$$\begin{aligned}
&\text{TERM 3} \\
&= \textstyle\prod_{is:\tau_{is}=k} \prod_{t=2}^{T} \left(z_{st} q_{ist} + (1 - z_{st})\right)^{y_{ist}} \left(1 - z_{st} q_{ist}\right)^{1-y_{ist}} \\
&= \underbrace{\prod_{is:\tau_{is}=k} \prod_{t=2}^{T} \left(\frac{z_{st} e^{\alpha_k + \beta z_{s,t-1} y_{is,t-1}}}{1+e^{\alpha_k + \beta z_{s,t-1} y_{is,t-1}}} + (1 - z_{st})\right)^{y_{ist}}}_{\text{TERM 3a}} \underbrace{\left(\frac{1 + (1-z_{st}) e^{\alpha_k + \beta z_{s,t-1} y_{is,t-1}}}{1+e^{\alpha_k + \beta z_{s,t-1} y_{is,t-1}}}\right)^{1-y_{ist}}}_{\text{TERM 3b}}
\end{aligned} \tag{7.42}$$

which comprises components

$$\begin{aligned}
\text{TERM 3a} &= \frac{\left(e^{\alpha_k+\beta}\right)^{\sum_{is} \sum_{t=2}^{T} I\left[\tau_{is}=k, z_{st}=1, z_{s,t-1}=1, y_{is,t-1}=1, y_{ist}=1\right]}}{\left(1+e^{\alpha_k+\beta}\right)^{\sum_{is} \sum_{t=2}^{T} I\left[\tau_{is}=k, z_{s,t-1}=1, y_{is,t-1}=1, y_{ist}=1\right]}} \\
&\quad \times \frac{\left(e^{\alpha_k}\right)^{\sum_{is} \sum_{t=2}^{T} I\left[\tau_{is}=k, z_{st}=1, y_{ist}=1, (z_{s,t-1}=0 \mid y_{is,t-1}=0)\right]}}{\left(1+e^{\alpha_k}\right)^{\sum_{is} \sum_{t=2}^{T} I\left[\tau_{is}=k, y_{ist}=1, (z_{s,t-1}=0 \mid y_{is,t-1}=0)\right]}}
\end{aligned} \tag{7.43}$$

and

$$\begin{aligned}
\text{TERM 3b} &= \frac{\left(1+e^{\alpha_k+\beta}\right)^{\sum_{is} \sum_{t=2}^{T} I\left[\tau_{is}=k, z_{st}=0, z_{s,t-1}=1, y_{is,t-1}=1, y_{ist}=0\right]}}{\left(1+e^{\alpha_k+\beta}\right)^{\sum_{is} \sum_{t=2}^{T} I\left[\tau_{is}=k, z_{s,t-1}=1, y_{is,t-1}=1, y_{ist}=0\right]}} \\
&\quad \times \frac{\left(1+e^{\alpha_k}\right)^{\sum_{is} \sum_{t=2}^{T} I\left[\tau_{is}=k, z_{st}=1, y_{ist}=0, (z_{s,t-1}=0 \mid y_{is,t-1}=0)\right]}}{\left(1+e^{\alpha_k}\right)^{\sum_{is} \sum_{t=2}^{T} I\left[\tau_{is}=k, y_{ist}=0, (z_{s,t-1}=0 \mid y_{is,t-1}=0)\right]}}
\end{aligned} \tag{7.44}$$

and hence

$$\text{TERM 3} = \frac{\left(e^{\alpha_k+\beta}\right)^{\Sigma_3}}{\left(1+e^{\alpha_k+\beta}\right)^{\Sigma_5}} \frac{\left(e^{\alpha_k}\right)^{\Sigma_4}}{\left(1+e^{\alpha_k}\right)^{\Sigma_6}} \tag{7.45}$$

where

$$
\begin{aligned}
\Sigma_3 = \quad & \textstyle\sum_{is}\sum_{t=2}^{T} \quad I[\tau_{is}=k,\ z_{st}=1,\ y_{ist}=1,\ z_{s,t-1}=1,\ y_{is,t-1}=1 \quad] \\
\\
\Sigma_4 = \quad & \textstyle\sum_{is}\sum_{t=2}^{T} \quad I[\tau_{is}=k,\ z_{st}=1,\ y_{ist}=1,\ z_{s,t-1}=0 \text{ or } y_{is,t-1}=0 \quad] \\
= \quad & \textstyle\sum_{is}\sum_{t=2}^{T} \quad I[\tau_{is}=k,\ z_{st}=1 \quad y_{ist}=1 \quad] - \Sigma_3 \\
\\
\Sigma_5 = \quad & \textstyle\sum_{is}\sum_{t=2}^{T} \quad \{I[\tau_{is}=k,\ z_{st}=1 \quad y_{ist}=1,\ z_{s,t-1}=1,\ y_{is,t-1}=1 \quad] \\
& \qquad\qquad -I[\tau_{is}=k,\ z_{st}=0,\ y_{ist}=0,\ z_{s,t-1}=1,\ y_{is,t-1}=1 \quad] \\
& \qquad\qquad +I[\tau_{is}=k,\ \qquad y_{ist}=0,\ z_{s,t-1}=1,\ y_{is,t-1}=1 \quad]\} \\
= \quad & \textstyle\sum_{is}\sum_{t=2}^{T} \quad \{I[\tau_{is}=k,\ z_{st}=1 \quad y_{ist}=1,\ z_{s,t-1}=1,\ y_{is,t-1}=1 \quad] \\
& \qquad\qquad +I[\tau_{is}=k,\ z_{st}=1,\ y_{ist}=0,\ z_{s,t-1}=1,\ y_{is,t-1}=1 \quad]\} \\
= \quad & \textstyle\sum_{is}\sum_{t=2}^{T} \quad I[\tau_{is}=k,\ z_{st}=1 \qquad\qquad z_{s,t-1}=1,\ y_{is,t-1}=1 \quad] \\
\\
\Sigma_6 = \quad & \textstyle\sum_{is}\sum_{t=2}^{T} \quad \{I[\tau_{is}=k,\ z_{st}=1 \quad y_{ist}=1,\ z_{s,t-1}=0 \text{ or } y_{is,t-1}=0 \quad] \\
& \qquad\qquad -I[\tau_{is}=k,\ z_{st}=0,\ y_{ist}=0,\ z_{s,t-1}=0 \text{ or } y_{is,t-1}=0 \quad] \\
& \qquad\qquad +I[\tau_{is}=k,\ \qquad y_{ist}=0,\ z_{s,t-1}=0 \text{ or } y_{is,t-1}=0 \quad]\} \\
= \quad & \textstyle\sum_{is}\sum_{t=2}^{T} \quad \{I[\tau_{is}=k,\ z_{st}=1 \quad y_{ist}=1,\ z_{s,t-1}=0 \text{ or } y_{is,t-1}=0 \quad] \\
& \qquad\qquad +I[\tau_{is}=k,\ z_{st}=1,\ y_{ist}=0,\ z_{s,t-1}=0 \text{ or } y_{is,t-1}=0 \quad]\} \\
= \quad & \textstyle\sum_{is}\sum_{t=2}^{T} \quad \{I[\tau_{is}=k \quad z_{st}=1 \qquad\qquad z_{s,t-1}=0 \text{ or } y_{is,t-1}=0 \quad]\} \\
= \quad & \textstyle\sum_{is}\sum_{t=2}^{T} \quad I[\tau_{is}=k \quad z_{st}=1 \quad\qquad\qquad\qquad\qquad\qquad] - \Sigma_5
\end{aligned}
\tag{7.46}
$$

The posterior ratio (used in Metropolis-Hastings style update) is

$$
\frac{p(\alpha_k^* \mid \ldots)}{p(\alpha_k \mid \ldots)} = e^{\Sigma_7(\alpha_k^* - \alpha_k)} \left(\frac{1+e^{\alpha_k}}{1+e^{\alpha_k^*}}\right)^{\Sigma_8} \left(\frac{1+e^{\alpha_k+\beta}}{1+e^{\alpha_k^*+\beta}}\right)^{\Sigma_5} \tag{7.47}
$$

where the exponents are

$$
\begin{aligned}
\Sigma_7 = \quad & \Sigma_1 + \Sigma_3 + \Sigma_4 \\
= \quad & \textstyle\sum_{is}\left\{I[\tau_{is}=k, z_{s1}=1, y_{is1}=1] + \sum_{t=2}^{T} I[\tau_{is}=k, z_{st}=1, y_{ist}=1]\right\} \\
= \quad & \textstyle\sum_{is}\sum_{t=1}^{T} I[\tau_{is}=k, z_{st}=1, y_{ist}=1] \\
\\
\Sigma_8 = \quad & \Sigma_2 + \Sigma_6 \\
= \quad & \textstyle\sum_{is}\{I[\tau_{is}=k, z_{s1}=1] \\
& \textstyle + \sum_{t=2}^{T} I[\tau_{is}=k, z_{st}=1] - I[\tau_{is}=k, z_{st}=1, z_{s,t-1}=1, y_{is,t-1}=1]\Big\} \\
= \quad & \textstyle\sum_{is}\sum_{t=1}^{T} I[\tau_{is}=k, z_{st}=1] \\
& \textstyle - \sum_{is}\sum_{t=2}^{T} I[\tau_{is}=k, z_{st}=1, z_{s,t-1}=1, y_{is,t-1}=1]
\end{aligned}
\tag{7.48}
$$

and for convenience recall that from equation (7.46)

$$\Sigma_5 = \sum_{is} \sum_{t=2}^{T} I[\tau_{is} = k, z_{st} = 1, z_{s,t-1} = 1, y_{is,t-1} = 1] \tag{7.49}$$

so

$$\Sigma_8 = \sum_{is} \sum_{t=2}^{T} I[\tau_{is} = k, z_{st} = 1] - \Sigma_5 \tag{7.50}$$

Marginal Posterior for unknown Z_{st}

The posterior for imputing missing values of Z has precisely the same structure as for EXTENSION I

$$p(z \mid \ldots) \propto p(y \mid q, z)\, p(z \mid \theta)$$

The difference is the form of the likelihood. Thus

$$\begin{aligned} p(z \mid \ldots) \propto \; & p(z \mid \theta) \prod_{is} p(y_{is1} \mid q_{is1}, z_{s1}) \prod_{t=2}^{T} p(y_{ist} \mid q_{ist}, z_{st}, y_{is,t-1}, z_{s,t-1}) \\ = \; & p(z \mid \theta) \prod_{is} \frac{(z_{s1} e^{\alpha \tau_{is1}} + (1 - z_{s1}))^{y_{is1}} \left(1 + (1 - z_{s1}) e^{\alpha \tau_{is1}}\right)^{1 - y_{is1}}}{1 + e^{\alpha \tau_{is1}}} \\ & \times \prod_{t=2}^{T} \frac{\left(z_{st} e^{\alpha \tau_{ist} + \beta y_{is,t-1} z_{s,t-1}} + (1 - z_{st})\right)^{y_{ist}} \left(1 + (1 - z_{st}) e^{\alpha \tau_{ist} + \beta y_{is,t-1} z_{s,t-1}}\right)^{1 - y_{ist}}}{1 + e^{\alpha \tau_{ist} + \beta y_{is,t-1} z_{s,t-1}}} \end{aligned} \tag{7.51}$$

In the case where $t = 1$ consider the subcases for each site s

Case $z_{s1} = 1$:

$$\begin{aligned} p(z_{s1} = 1 \mid \ldots) \propto \; & p(z \mid \theta) \prod_{i=1}^{I} \frac{(e^{\alpha \tau_{is1}})^{y_{is1}}}{1 + e^{\alpha \tau_{is1}}} \\ & \times \frac{(z_{s2} e^{\alpha \tau_{is2} + \beta y_{is1}} + (1 - z_{s2}))^{y_{is2}} (1 + (1 - z_{s2}) e^{\alpha \tau_{is2} + \beta y_{is1}})^{1 - y_{is2}}}{1 + e^{\alpha \tau_{is2} + \beta y_{is1}}} \end{aligned}$$

Case $z_{s1} = 0$:

$$\begin{aligned} p(z_{s1} = 0 \mid \ldots) \propto \; & p(z \mid \theta) \prod_{i=1}^{I} \frac{(1 + e^{\alpha \tau_{is1}})^{1 - y_{is1}}}{1 + e^{\alpha \tau_{is1}}} \\ & \times \frac{(z_{s2} e^{\alpha \tau_{is2}} + (1 - z_{s2}))^{y_{is2}} (1 + (1 - z_{s2}) e^{\alpha \tau_{is2}})^{1 - y_{is2}}}{1 + e^{\alpha \tau_{is2}}} \end{aligned}$$

So the Odds ratio is:

$$\begin{aligned} \frac{p(z_{s1} = 1 \mid \ldots)}{p(z_{s1} = 0 \mid \ldots)} = \; & \frac{p(z_{s1} = 1 \mid z_{-s1}, \theta)}{p(z_{s1} = 0 \mid z_{-s1}, \theta)} \\ & \times \prod_{i=1}^{I} \frac{(e^{\alpha \tau_{is1}})^{y_{is1}}}{(1 + e^{\alpha \tau_{is1}})^{1 - y_{is1}}} \left(z_{s2} e^{\beta y_{is1}} + (1 - z_{s2})\right)^{y_{is2}} \\ & \times \left(\frac{(1 + (1 - z_{s2}) e^{\alpha \tau_{is2} + \beta y_{is1}})}{(1 + (1 - z_{s2}) e^{\alpha \tau_{is2}})} \right)^{1 - y_{is2}} \frac{1 + e^{\alpha \tau_{is2}}}{1 + e^{\alpha \tau_{is2} + \beta y_{is1}}} \end{aligned}$$

When $t > 1$ the marginal posterior distribution of z_{st} is

$$p(z_{st} \mid \ldots) \propto p(z \mid \theta) \times \prod_{i=1}^{I} \frac{(z_{st} e^{\alpha\tau_{ist}+\beta z_{s,t-1}y_{is,t-1}} + (1 - z_{st}))^{y_{ist}} (1 + (1 - z_{st}) e^{\alpha\tau_{ist}+\beta z_{s,t-1}y_{is,t-1}})^{1-y_{ist}}}{1 + e^{\alpha\tau_{ist}+\beta z_{s,t-1}y_{is,t-1}}} \times \frac{(z_{s,t+1} e^{\alpha\tau_{is,t+1}+\beta z_{st}y_{ist}} + (1 - z_{s,t+1}))^{y_{is,t+1}} (1 + (1 - z_{s,t+1}) e^{\alpha\tau_{is,t+1}+\beta z_{st}y_{ist}})^{1-y_{is,t+1}}}{1 + e^{\alpha\tau_{is,t+1}+\beta z_{st}y_{ist}}} \tag{7.52}$$

Then for $t \geq 2$

Case $z_{st} = 0$

$$p(z_{st} = 0 \mid \ldots) \propto p(z \mid \theta) \prod_{i=1}^{I} \frac{(1 + e^{\alpha\tau_{ist}+\beta z_{s,t-1}y_{is,t-1}})^{1-y_{ist}}}{(1 + e^{\alpha\tau_{ist}+\beta z_{s,t-1}y_{is,t-1}})} \times \frac{(z_{s,t+1} e^{\alpha\tau_{is,t+1}} + (1 - z_{s,t+1}))^{y_{is,t+1}} (1 + (1 - z_{s,t+1}) e^{\alpha\tau_{is,t+1}})^{1-y_{is,t+1}}}{1 + e^{\alpha\tau_{is,t+1}}} \tag{7.53}$$

Case $z_{st} = 1$

$$p(z_{st} = 0 \mid \ldots) \propto p(z \mid \theta) \prod_{i=1}^{I} \frac{(e^{\alpha\tau_{ist}+\beta z_{s,t-1}y_{is,t-1}})^{y_{ist}}}{(1 + e^{\alpha\tau_{ist}+\beta z_{s,t-1}y_{is,t-1}})} \times \frac{(z_{s,t+1} e^{\alpha\tau_{is,t+1}+\beta y_{ist}} + (1 - zs, t + 1))^{y_{is,t+1}} (1 + (1 - z_{s,t+1}) e^{\alpha\tau_{is,t+1}+\beta y_{ist}})^{1-y_{is,t+1}}}{1 + e^{\alpha\tau_{is,t+1}+\beta y_{ist}}} \tag{7.54}$$

and the Odds ratio is

$$\begin{aligned} \frac{p(z_{st} = 1 \mid \ldots)}{p(z_{st} = 0 \mid \ldots)} = & \frac{p(z_{st} = 1 \mid z_{-st}, \theta)}{p(z_{st} = 0 \mid z_{-st}, \theta)} \times \\ & \prod_{i=1}^{I} \left[\frac{\left(e^{\alpha\tau_{ist}+\beta z_{s,t-1}y_{is,t-1}}\right)^{y_{ist}}}{\left(1 + e^{\alpha\tau_{ist}+\beta z_{s,t-1}y_{is,t-1}}\right)^{1-y_{ist}}} \frac{(1 + e^{\alpha\tau_{is,t+1}})}{\left(1 + e^{\alpha\tau_{is,t+1}+\beta y_{ist}}\right)} \right. \\ & \times \left(z_{s,t+1} e^{\beta y_{ist}} + (1 - z_{s,t+1})\right)^{y_{is,t+1}} \\ & \times \left. \left(\frac{1 + (1 - z_{s,t+1}) e^{\alpha\tau_{is,t+1}+\beta y_{ist}}}{1 + (1 - z_{s,t+1}) e^{\alpha\tau_{is,t+1}}} \right)^{1-y_{is,t+1}} \right] \end{aligned} \tag{7.55}$$

For case $t = 1$ we observe that the only difference is in the first term where $z_{s0} = y_{is0} = 0$.

Note that the prior ratio, which enters into the odds ratio for the posterior distribution of z_{st} evaluates to

$$\frac{p(z_{st} = 1 \mid z_{-st}, \theta)}{p(z_{st} = 0 \mid z_{-st}, \theta)} = \exp\{\theta_0 + \theta_1(z_{s-1,t} + z_{s+1,t}) + \theta_2(z_{s,t-1} + z_{s,t+1})\} \tag{7.56}$$

Now there are 2^5 permutations of the parameters $z_{s,t-1}$, $y_{is,t-1}$, y_{ist}, $y_{is,t+1}$, $z_{s,t+1}$. The large product term in the odds ratio above contained within square brackets [] in equation (7.55) can be simplified by considering groups of these permutations which lead to the same value

of the odds ratio. For example in the case where all five quantities evaluate to zero, only the first fraction in the expression remains; all others evaluate to one. These equations are simplified in Table 7.57 as follows assuming i and s held constant throughout the table, and dropping the subscripts for i and s for simplicity. The column labelled #Perm counts the number of permutations on the last line for which the odds ratio has the value given in the first column.

Product in equation (7.55)	**#Perm**	z_{t-1}	y_{t-1}	y_t	y_{t+1}	z_{t+1}
Disallow combinations where	(14)	0	1	*	*	*
$z = 0$ and $y = 1$		*	*	*	1	0
$\frac{1}{1+e^{\alpha_{\tau_t}}}$	(6)	*	0	0	$(* - *)$	
$\frac{1}{1+e^{\alpha_{\tau_t}+\beta}}$	(3)	1	1	0	$(* - *)$	
$e^{\alpha_{\tau_t}}$	(2)	*	0	1	0	0
$e^{\alpha_{\tau_t}+\beta}$	(1)	1	1	1	0	0
$e^{\tau_t}\frac{(1+e^{\alpha_{\tau_t}})}{(1+e^{\alpha_{\tau_t}+\beta})}$	(2)	*	0	1	0	1
$e^{\tau_t+\beta}\frac{(1+e^{\alpha_{\tau_t}})}{(1+e^{\alpha_{\tau_t}+\beta})}$	(3)	1	1	1	0	1
		*	0	1	1	1
$e^{\tau_t+2\beta}\frac{(1+e^{\alpha_{\tau_t}})}{(1+e^{\alpha_{\tau_t}+\beta})}$	(1)	1	1	1	1	1
Total	(32)					

(7.57)

Here $*$ is a wildcard matching either of the possibilities $\{0, 1\}$ and $(* - *)$ is a wildcard matching any of the allowable possibilities $\{(0,0), (0,1), (1,1)\}$.

Marginal Posterior for θ

However the posterior distribution for Presence dependence θ does change with the extensions to the hierarchical model due to the additional level incorporated to describe its distribution $p(\theta)$. So the posterior for θ depends only on the underlying presence model and its prior distribution.

$$p(\theta \mid \ldots) \propto p(z \mid \theta) p(\theta). \tag{7.58}$$

Marginal Posterior for β

The posterior ratio used in Metropolis-Hastings style updates of β only involves likelihood ratios and prior ratios:

$$\frac{p(\beta^* \mid \ldots)}{p(\beta \mid \ldots)} = \frac{p(\beta^*)}{p(\beta)} \prod_{ist} \frac{p(y_{ist} \mid q^*_{ist}, z)}{p(y_{ist} \mid q_{ist}, z)} \frac{p(q^*_{ist})}{p(q_{ist})} \tag{7.59}$$

where q^*_{ist} is defined by the usual expression for the linear predictor given equation (7.27) with β^* substituted for β.

7.3.9 Computation via MCMC

As for EXTENSION I, a systematic or random sampling régime can be used for updating all random variable components. Also as for EXTENSION I, the type of sampler used for updates is tailored to the posterior distribution of the component. Again, the only parameter having a posterior easy to sample from is the autologistic variable Z_{st}. All other random variables—β, α_k, and θ—are easier to update using Metropolis-Hastings updates which only require ratios of the posterior distributions.

We now give details of the MCMC computations for each random variable component separately.

7.3.10 MCMC sampler for β

An appropriate starting value for β in this case is the central value of its prior distribution

$$\beta^{(0)} = 0.0.$$

Uniform distributions and either Normal or Uniform distributions with smaller variance were considered, but did not provide adequate movement within the chains (acceptance rates were too low or too high.) Another justification for the Normal distribution is that it is a common choice for the distribution of parameters within hierarchical linear models (*e.g.* Gelman et al. (1995)).

A Metropolis-Hastings update requires a proposal distribution. A Normal distribution having the same variance and truncated to the same interval as the prior ($[\beta_{\mathsf{L}}, \beta_{\mathsf{U}}]$) is again suitable. Although having a truncated prior will automatically ensure that the product of the prior and proposal ratios is truncated, it is still computationally more efficient to ensure that only those values are proposed that fall within the truncated interval. Instead of being centred near zero, the proposal distribution is centred at the previous estimate β. That is,

$$\beta^* \mid \beta \sim \mathrm{N}(\beta, \sigma^2_\beta) \cap [\beta^*_{\mathsf{L}}, \beta^*_{\mathsf{U}}].$$

and so the proposal ratio is given in a previous Section 4.4.2.

The posterior ratio used in Metropolis-Hastings style updates of β only involves likelihood ratios and prior ratios:

$$\frac{p(\beta^* \mid \ldots)}{p(\beta \mid \ldots)} = \prod_{ist} \frac{p(y_{ist} \mid q^*_{ist}), z}{p(y_{ist} \mid q_{ist}, z)} \frac{p(q^*)}{p(q)} \tag{7.60}$$

MCMC sampler for α_k

The posterior distribution for α_k is not easy to sample from, although a ratio of posterior distribution values is easy to evaluate. We therefore use the Metropolis-Hastings design of an MCMC sampler to update each component α_k.

This requires a proposal distribution. Some possibilities are a Normal distribution in which case the ratio of proposals is 1, or a one-parameter logistic distribution centred on the old value.

MCMC sampler for Z_{st}

See the discussion for the MCMC sampler for z_{st} in Extension I given in Section 7.2.9. The posterior distribution and the posterior ratio for z_{st} is given in equations (7.16)–(7.19). Truncating both $V_{st}(z)$ and θ from three components to their first two components only adapts these equations to this Extension II.

7.3.11 MCMC sampler for θ

In practice, a finite but large number of θ values will be compared. Therefore a Metropolis-Hastings update is more efficient than a Gibbs sampling approach. The Metropolis-Hastings ratio of posteriors, new proposed (*) compared to old, involves the ratio of normalization constants, and is given by

$$\frac{p(\theta^* \mid \ldots)}{p(\theta \mid \ldots)} = \exp\{(\theta^* - \theta)^\top V(z)\} \frac{c(\theta)}{c(\theta^*)} \frac{p(\theta^*)}{p(\theta)} \tag{7.61}$$

The real challenge here with the θ update is the unknown normalization constant ratio $\frac{c(\theta)}{c(\theta^*)}$. This issue is precisely the same as that encountered for Extension I of the model, and is discussed in Section 7.2.10.

New θ^* values can be generated from a discrete uniform proposal distribution centred on the current value of θ in 2-dimensional space, and extending up to two "jumps" in any direction ($h_l = 2,\ l = 0, 1$).

Starting values for θ were chosen from the discrete uniform distribution over the range of $\{\theta_m\}$ selected for the experiment.

7.4 Results: overview

We shall apply the EXTENSION I model to the *dingo* case study to evaluate its effectiveness.

Although IMCS and RLR both contributed greatly to permitting computation required for modelling approaches EXTENSIONS I AND II, they do constrain the analysis since they only provide estimates of ratios of normalization constants. Thus the success of modelling depends on a good choice of the discrete choice of values θ_m from parameter space Θ. We begin with a small-scale low-resolution lattice in the first stage, experiment 1 (this section), and then "zoom in" on the most successful section of the lattice in the second stage, Experiment 2 (Section 7.6). In practice, we require the resolution of the Θ lattice describing spatio-temporal dependence to be sufficiently precise for scientific description of the underlying processes. This sequential approach to design and analysis is just the usual scientific method where the 'experiments' are computational rather than tangible.

Thus we begin by considering one base permutation of starting values or proposal distributions and prior distributions. Results are analyzed in depth for this base permutation. Some sensitivity analysis is conducted by changing some starting assumptions and measuring the impact on results.

In order to refine the MCMC algorithm, the chain burnin, length, thinness were varied until convergence was attained. Results from Chapter 5 were referred to in order to select appropriate prior and proposal distributions and parameters. Results obtained for other choices were similar or worse and are not reported in detail here.

Final distributional choices for the first stage, Experiment 1 are tabulated in Table C.1. Here NN refers to a nearest neighbour model on a discrete d-dimensional space of the specified order. A nearest neighbour is defined as being any point in the space that can be reached within steps of the specified order, *e.g.* for order 2, all points which are either 1 or 2 steps away from the point in any direction parallel to the axes. Truncations of proposal and prior distributions are desired mainly for computational reasons. Final distributional choices for the second stage, Experiment 2 are tabulated in Table C.2.

More extensive and advanced MCMC diagnostics are used for assessing convergence and efficacy of MCMC simulation algorithms in this chapter compared to preliminary work on the Pilot experiment in Chapter 5. A detailed explanation of how the MCMC diagnostics described in Section 4.4.7 have been applied to simulations is given for analysis of this particular experiment only. Later experiments focus more on examination of the posterior distributions.

7.5 Results: EXTENSION I, Experiment 1

The MCMC chains were obtained from a total of 2 million (2,000,000) runs after a burnin of 200,000. A thinning factor of 100 was used, resulting in 20,000 samples used to derive simulations from the posterior distributions. The burnin and thinning factors were selected in order to ensure convergence, according to the diagnostics: Geweke's batch statistic; Raftery-Lewis' quantile estimation statistic; and the Heidelberg-Welch stationarity tests. Gelman-Rubin multiple chain tests were not used to formally assess convergence from various starting points. Instead, informal inspection of results from various starting values, combined with initial results in Section 5.5.5 indicated that starting values were good choices. This avoided high disk space requirements of storing the multiple chains in order to implement the Gelman-Rubin test. Runs of 20,000 were initially considered with zero burnin and no thinning. However all diagnostics were failed by at least one parameter. In particular the autocorrelation estimates and Geweke plots indicated extreme "stickiness" in the chains. Thinning by 100 eliminated most of this problem.

7.5.1 Posterior distribution of θ

The presence/absence process X is driven by the parameter θ which comprises three components. We inspect each of the three two-dimensional histograms—one for each pair of components—supplemented by their corresponding one-dimensional histograms. We see that the posterior distribution is not symmetric in θ components, but rather a combination of θ_1 and θ_2, indicating a correlation between spatial and temporal dependence.

Marginal posterior distributions of each θ component

We first inspect the marginal one-dimensional histograms, one for each component of θ. Figure 7.3 shows the marginal distribution of each of the components of theta, both as a histogram showing the distribution aggregated over time, and also as a time series showing the stability of the distribution over time. Note that the time series indicate that the overall distribution is the same over the entire time period, and therefore not providing any evidence contradicting that the marginal distributions of each of the θ components is in equilibrium.

For θ_0, even the least supported value, -1.7, in the marginal posterior distribution of prevalence was observed in nearly 15% of simulations. The most common value was -1.95 and was observed in over half the simulations. The other value -2.2 was observed in nearly one-third of simulations. This suggests that perhaps a wider and finer grid of prevalence values need be explored in more depth.

For θ_1, a value of 0.5 far outweighs all other values for spatial dependence (given a choice between 0, 0.5, 1.0 and 1.5.) The next most common value for spatial dependence is 0.0.

For θ_2, note that values of 0.5 and 1.0 for temporal dependence are given almost equal support (frequency 50% vs frequency 45%).

Conditional Posterior distributions

Fig 7.3 shows the relative posterior distribution of θ_0 conditional on each combination of the θ_1, θ_2 tuples. The distribution over prevalence definitely does depend on spatial dependence: for low spatial dependence, prevalence parameters closer to -1.7 are more common, whereas for higher spatial dependence, prevalence parameters tend to be closer to the -2.2 value on average. Inspection of the y-axes reveals that the three most frequently occurring combinations of spatial and temporal dependence tuples, *i.e.* (0,1), (0.5,0.5), (0.5,1). For each of these combinations, the most common prevalence parameter value is different. Values near -1.7 and -1.95 are more common for the first tuple, compared to -1.95 for the second and -2.2 for the third. Thus there appears to be a complex interaction between prevalence and spatial and temporal dependence.

Pairwise Posterior distributions

The top and left margins of figure 7.4 show the marginal posterior distributions of the spatial and temporal dependence parameters, θ_1 and θ_2 respectively. Comparison of these results suggest that temporal dependence has slightly more impact on presence/absence, being in predominantly in the range $0.5 - 1.0$, compared to spatial dependence, which lies mostly in the range of $0 - 0.5$.

The bottom right panel of figure 7.4 shows the marginal bivariate posterior distribution of both spatial and temporal dependence parameters. Here it is obvious that the three most common combinations of (θ_1, θ_2) are pairs $(0.5, 0.5)$, $(0.5, 1)$ and $(0, 1)$, in decreasing order of support. It is interesting that the posterior mode is located at an isotropic pair, that is spatial and temporal dependence are equal.

The modal value of the bivariate posterior distribution of spatial dependence and prevalence $(0.5, -1.95)$ coincides with the modal values of each of the marginal posterior distributions. Another almost as common modal value of the bivariate posterior was $(0.5, -2.2)$. These two common points share spatial dependence of 0.5. The next three most common values have low frequency and occur at spatial dependence and prevalence pairs of

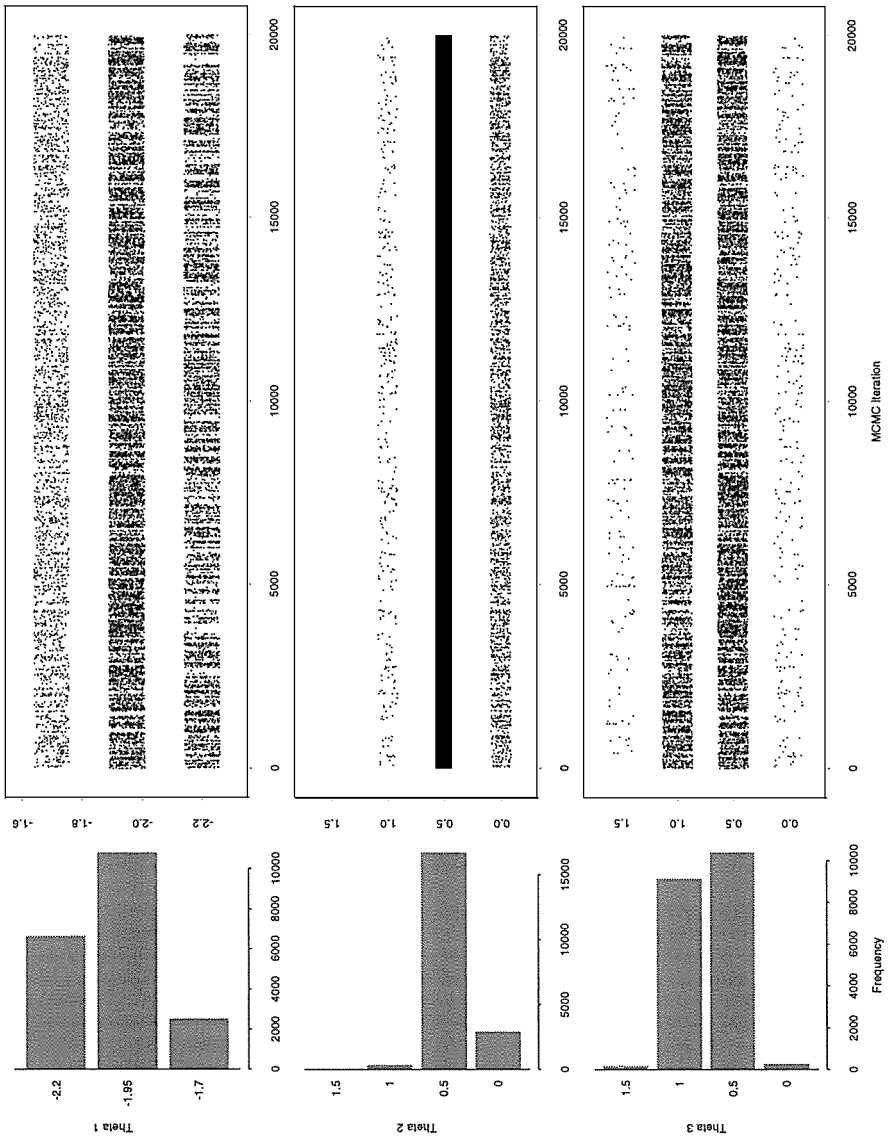


Figure 7.3: Extension I, Experiment 1: Univariate posterior distribution of each component of θ. Note that the y-axis is the values of θ_l for plots on the left and the right hand side.

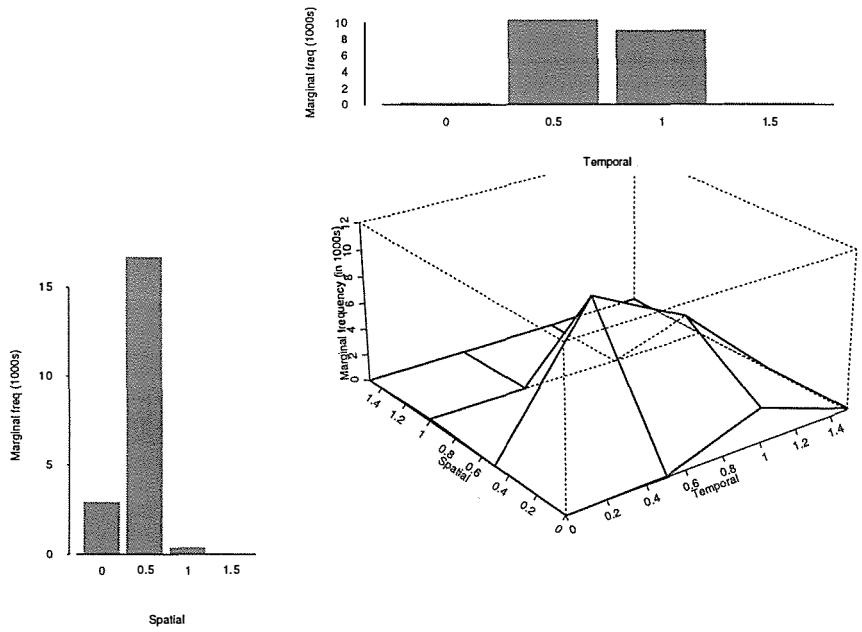


Figure 7.4: Extension I, Experiment 1: Bivariate posterior distribution of (θ_1, θ_2), two components of θ.

$(0, -1.95)$, $(0, -1.7)$, and $(0.5, -1.7)$. Almost no support is shown for values with spatial dependence of 1 or 1.5.

The marginal posterior distributions of temporal dependence and prevalence are focussed around values of $(0.5, 1.0)$ and $(-2.2, -1.95)$ respectively. Their bivariate posterior distribution achieves a modal value at the pair $(0.5, -1.95)$ which corresponds to the modal values of the univariate posterior distributions. The second almost as common value in the bivariate posterior is the pair $(1, -2.2)$. It is interesting that is not $(0.5, -2.2)$ nor $(1, -1.95)$. The next most common values have low frequency and are situated at $(0.5, -1.7)$, $(1, -1.7)$ and $(1, -1.95)$. None of the pairs corresponding to temporal dependence values 0 and 1.5 have support in the bivariate posterior distribution.

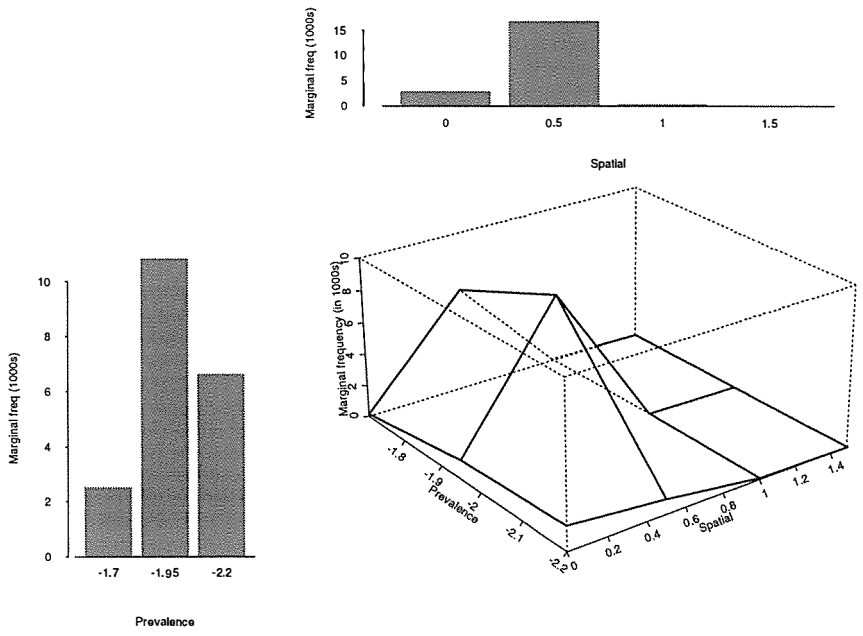


Figure 7.5: Extension I, Experiment 1: Bivariate posterior distribution of (θ_0, θ_1), two components of θ.

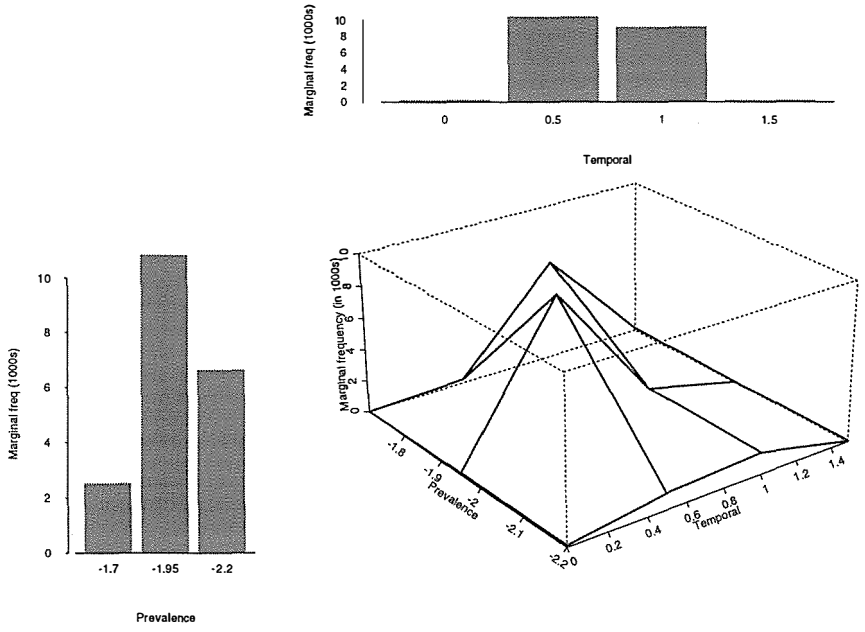


Figure 7.6: Extension I, Experiment 1: Bivariate posterior distribution of (θ_0, θ_2), two components of θ.

Posterior distribution of three-dimensional θ

In the full three-dimensional frequency plot shown in figure 7.7 the width and height of the rectangle are proportional to the frequency of the triplet of prevalence, spatial dependence and temporal dependence corresponding to each value of θ in the design space of the experiment. This approach emphasises the more frequent triplets so that they can be easily distinguished from the others.

Two points in the three-way histogram stand out. The most common point occurs for prevalence -1.95, spatial dependence 0.5 and temporal dependence 0.5. The next most common point is for prevalence -2.2, spatial dependence 0.5 and temporal dependence 1.

The more common points appear to occur within a "banana" shaped region concentrated at $(-2.2, 0.5, 1)$ and $(-1.95, 0.5, 0.5)$ and tapering off towards $(-1.7, 0, 1)$, $(-1.7, 0.5, 0.5)$ and $(-1.7, 1, 0)$. This suggests some correlation between the parameters. As prevalence parameter increases, the temporal dependence parameter decreases at a faster rate than the spatial dependence parameter. This is a well-known artifact of spatial models (Pettitt & Low Choy 1999, Weir & Pettitt 1999, Wolpert & Ickstadt 1998): a trade-off exists between the overall density and between perceived dependence between observations.

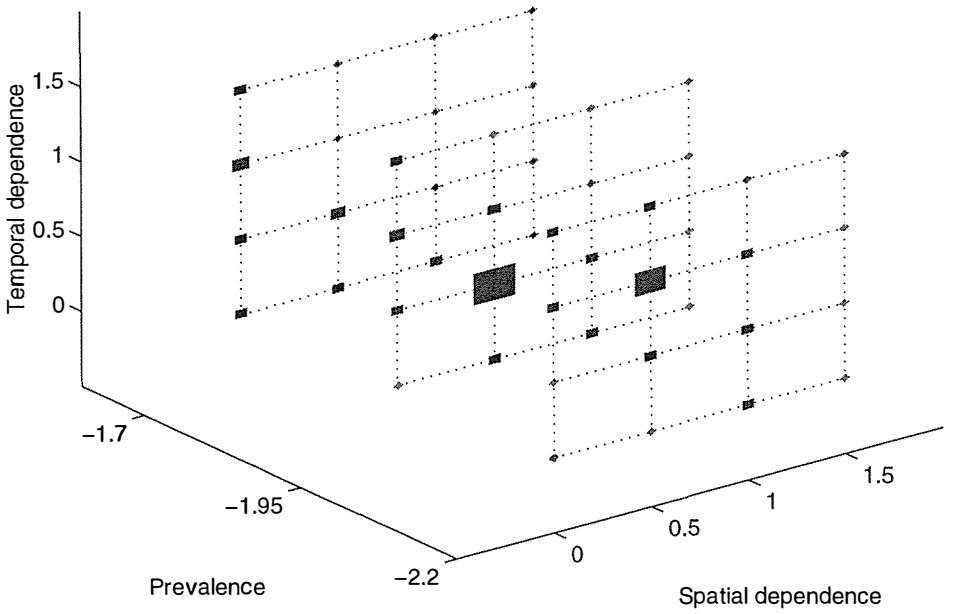


Figure 7.7: Extension I, Experiment 1: Three-dimensional histogram of the joint marginal posterior distribution of the three θ components

7.5.2 Proposal distribution of θ

Acceptance probabilities are important indicators of how "sticky" or dependent the MCMC chain is. In this case with a sparse discrete parameter space, they need to be interpreted carefully.

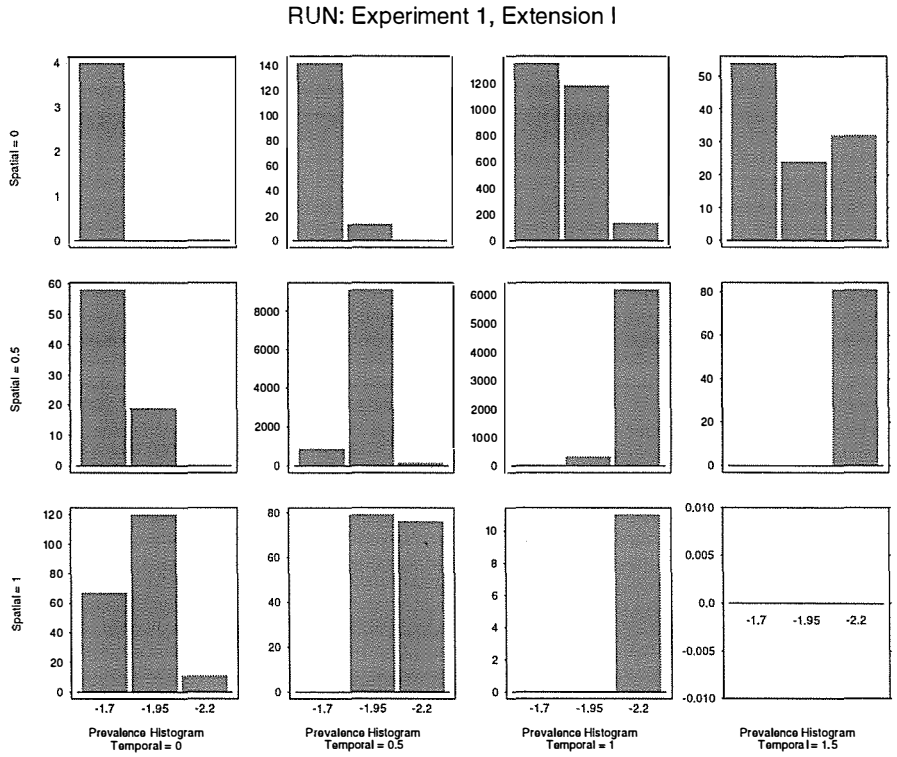


Figure 7.8: Extension I, Experiment 1: Plot of joint posterior distribution of θ for each combination of spatial parameter θ_1 (y-axis) and temporal parameter θ_2 (x-axis). Panel plots show frequency of each prevalence parameter θ_0. Note that the plot in the lower right corner has no values.

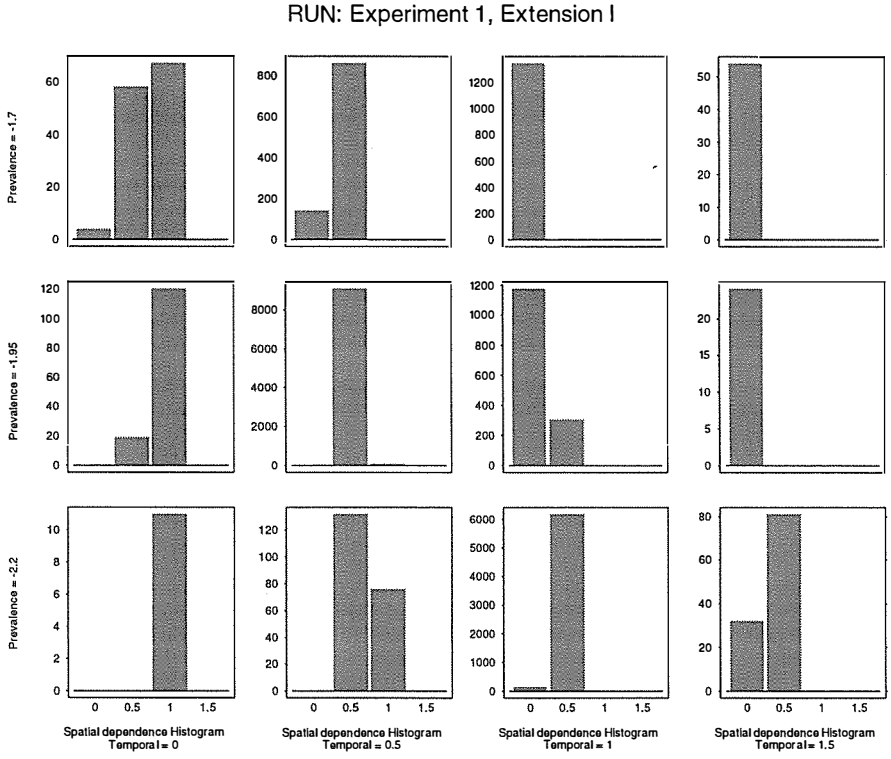


Figure 7.9: Extension I, Experiment 1: Plot of joint posterior distribution of θ for each combination of prevalence parameter θ_0 (y-axis) and temporal parameter θ_2 (x-axis). Panel plots show frequency of each spatial parameter θ_1.

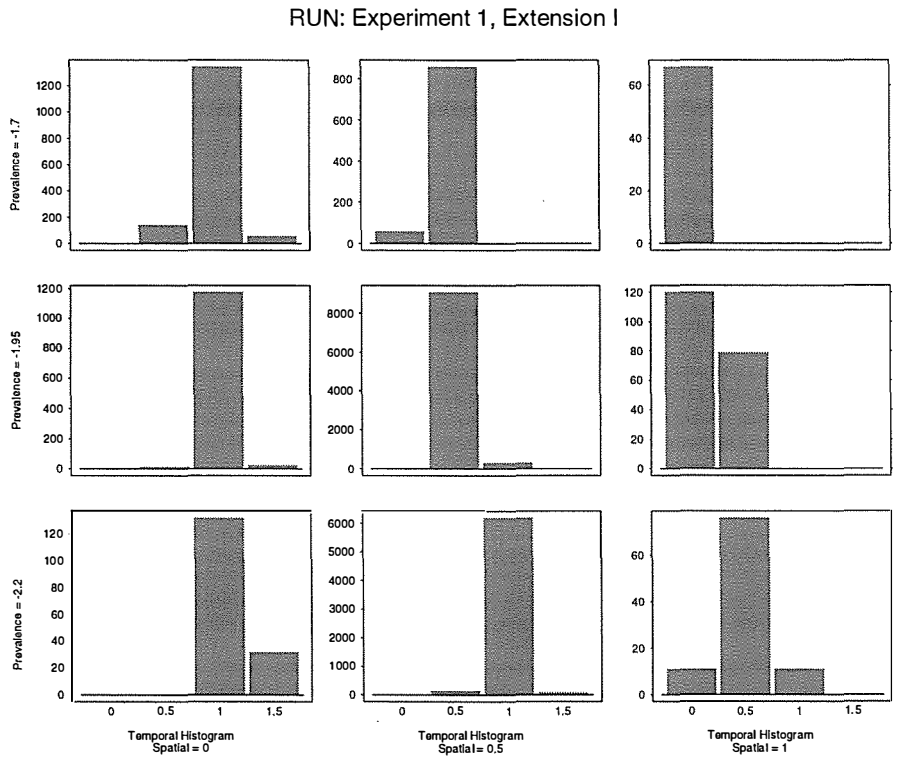


Figure 7.10: Extension I, Experiment 1: Plot of joint posterior distribution of θ for each combination of prevalence parameter θ_0 (y-axis) and spatial parameter θ_1 (x-axis). Panel plots show frequency of each temporal parameter θ_2.

Proposal rates indicate whether all portions of the parameter space had sufficient opportunity to be selected for updating. In Figure 7.11 it is obvious that parameter values, other than modal values identified above, were proposed often, and that the design space for Θ was completely explored. This is also a feature of the proposal distribution.

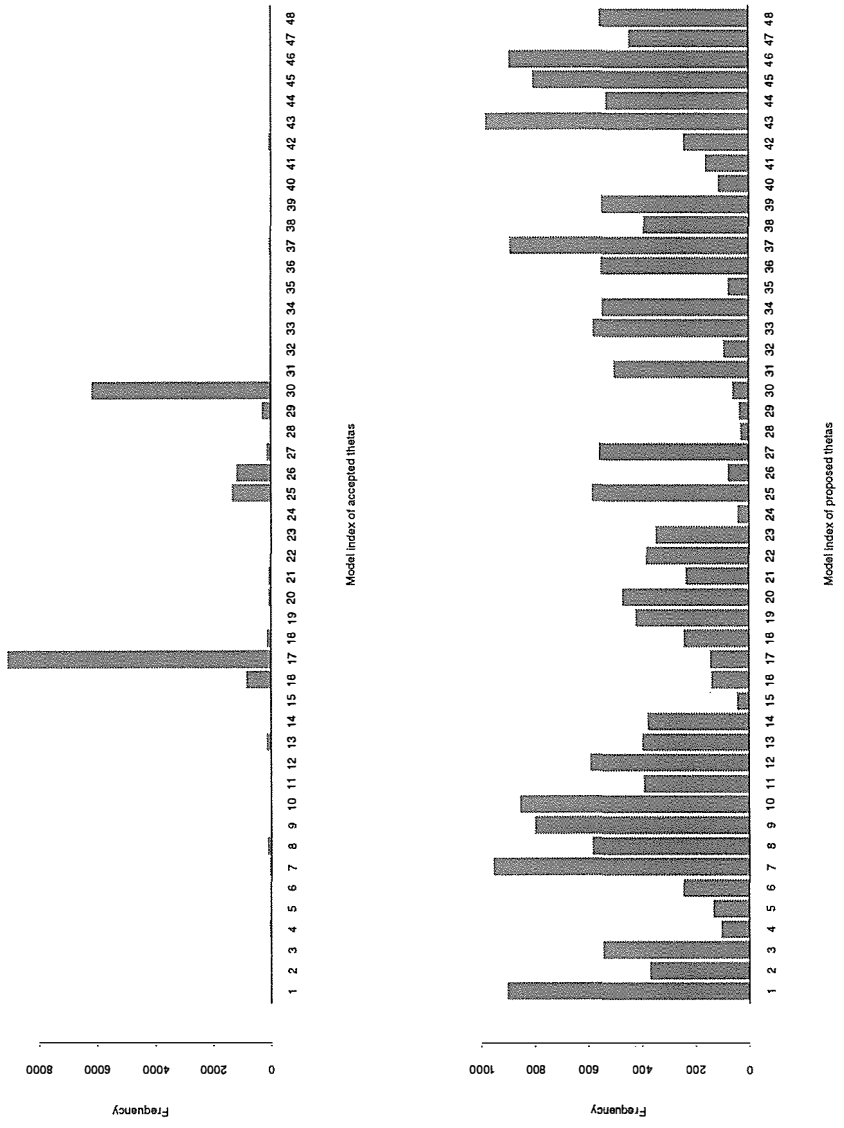


Figure 7.11: Extension I, Experiment 1: Marginal proposal distribution of each discrete $\theta_{(m)}$. Top figure represents accepted θ index m, and bottom figure contains suggested values θ index m'.

7.5.3 Posterior distribution of natural statistics of presence/absence

Corresponding to each of the three components of θ, is one of three components of the natural statistic $V(x)$.

The natural statistic estimates are reported for the entire space-time lattice. However a large proportion of sites on this lattice are unknown (corresponding to an inconclusive recorded non-visit $y_{st} = 0$). The contributions to the natural statistics of presence due to the observed portion of the presence/absence lattice are:

$$\begin{aligned} V_0^{(\text{known})}(x) &= 106 \\ V_1^{(\text{known})}(x) &= 21 \\ V_2^{(\text{known})}(x) &= 23 \end{aligned} \tag{7.62}$$

Descriptive statistics for the whole presence/absence lattice are given in Table 7.1. The natural statistic representing the total amount of presence (out of 945 spatio-temporal sites) is centred at 170 with a relatively tight standard error of 19. The lag 1 autocorrelation in the MCMC chain is quite small (0.39) indicating that the chain for presence is not too 'sticky' and need not be thinned more.

In figure 7.12 the compact trace plot depicts no consistent changes in the level of the posterior distribution of any of the three components. An expanded view of these trace plots, demonstrates that even at finer time resolution, the level is maintained consistently throughout the concentrations (since the lag 1 correlation was 0.39). This evidence is consistent with the conclusion that the MCMC chain has converged to equilibrium and that the values obtained are therefore representative of the posterior distribution.

Further diagnostics were investigated to determine whether this convergence was transitory or representative of equilibrium. These are documented in table 7.2. The Geweke test indicates convergence for each of the natural statistics. The Raftery-Lewis estimates of sample size agree that at most 5500 iterations in the chain are required in order to estimate the 0.025 and the 0.975 quantiles accurately. The Heidelberger-Welch stationarity and half-width tests indicate that the chains are all stationary.

7.5.4 Posterior distribution of effects of explanatory variables

For each of the unknown parameters in the linear predictor, an MCMC trace of the simulated posterior distribution is shown in figures 7.13 and 7.14. All effects are portrayed on the probability scale q_k of probability of chemical attraction for each of the $k = \tau_{ist}$ chemicals. Despite being probabilities all their marginal posterior distributions are approximately Normal, except for that relating to the lowest estimated probability q_2, which is skewed left towards low values, which is to be expected due to the boundary effect at 0. Stationarity of the mean is evident.

Table 7.3 gives more descriptive statistics of the posterior simulations. The most powerful probability of chemical attraction occurs for close contenders chemical A ($q_1 = 0.56$, $\text{sd}(q) = 0.077$) and chemical D ($q_4 = 0.57$, $\text{sd}(q) = 0.091$), with the latter having less precision. These two chemicals are followed in equal steps by chemicals 5, 6, 3 and 2 (in that order). This coincides with the orderings allocated in virtually all other analyses of this data. See Chapter 3.

There is sufficient serial correlation to have a small impact on the independent sample estimate of standard error. The naive, time series and batch estimates of standard errors in

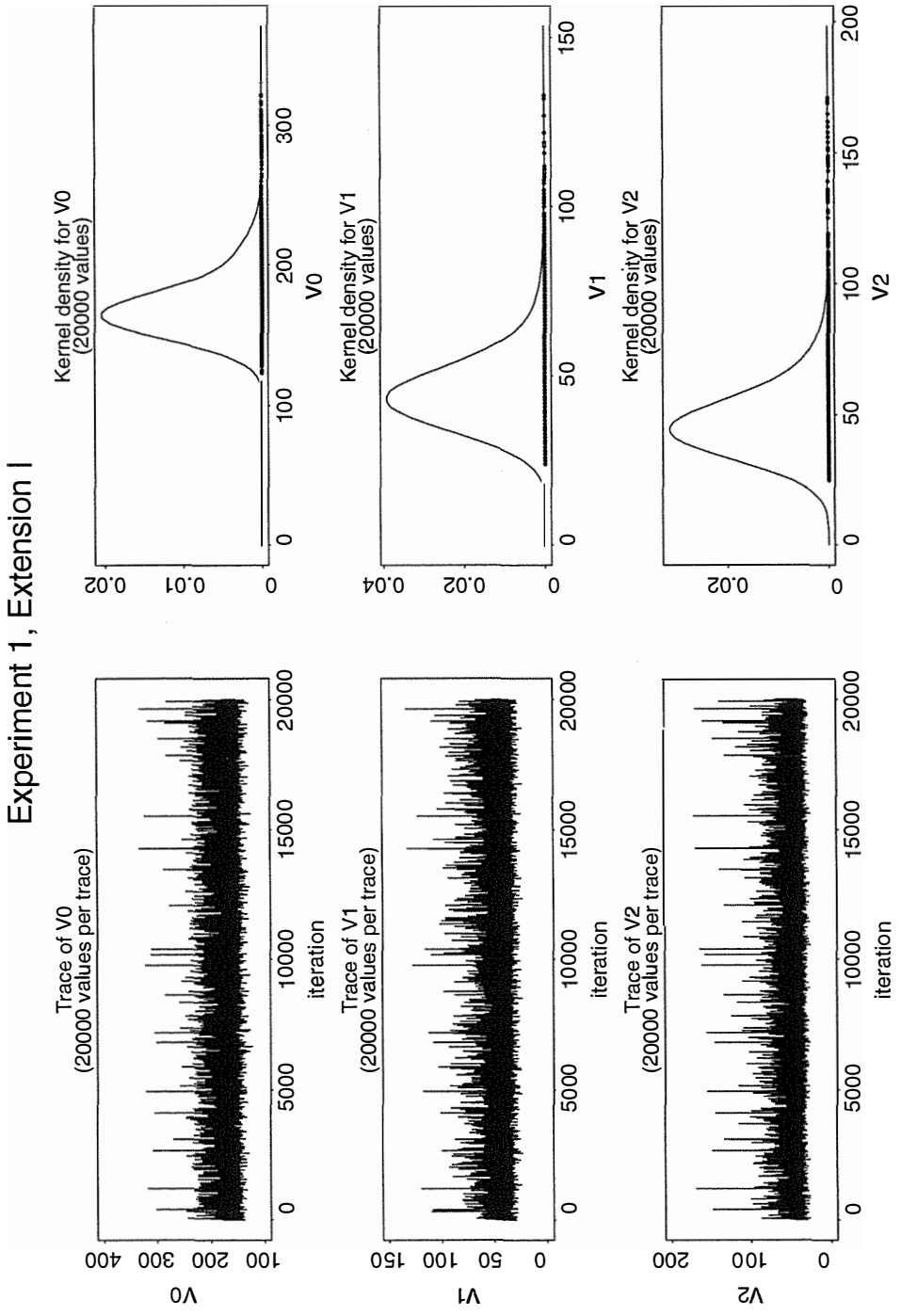


Figure 7.12: Extension I, Experiment 1: MCMC Trace of the posterior distribution of natural statistics of presence/absence $V(x)$

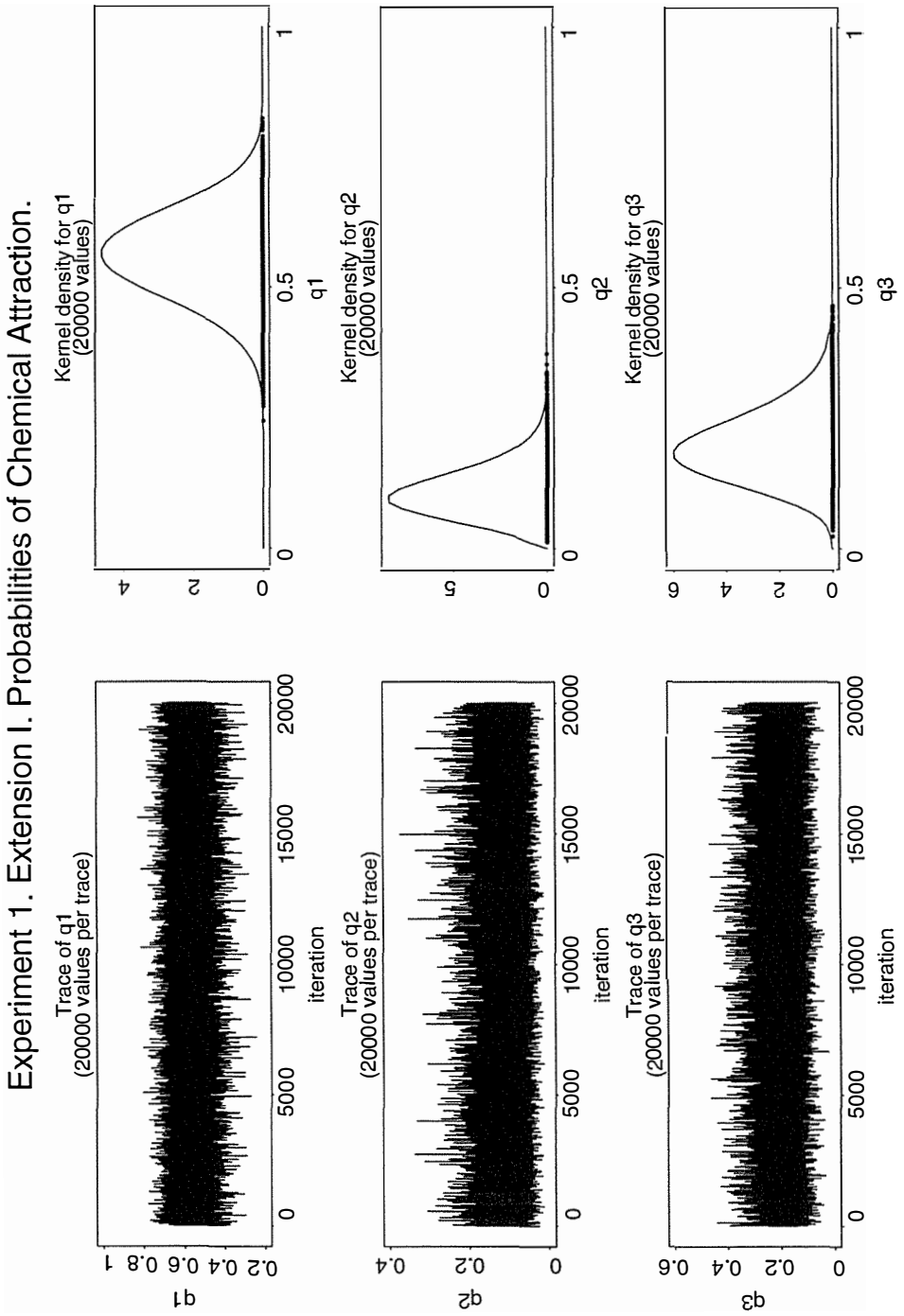


Figure 7.13: Extension I, Experiment 1: MCMC Trace of the posterior distribution of effects of explanatory variables q_1, q_2, q_3.

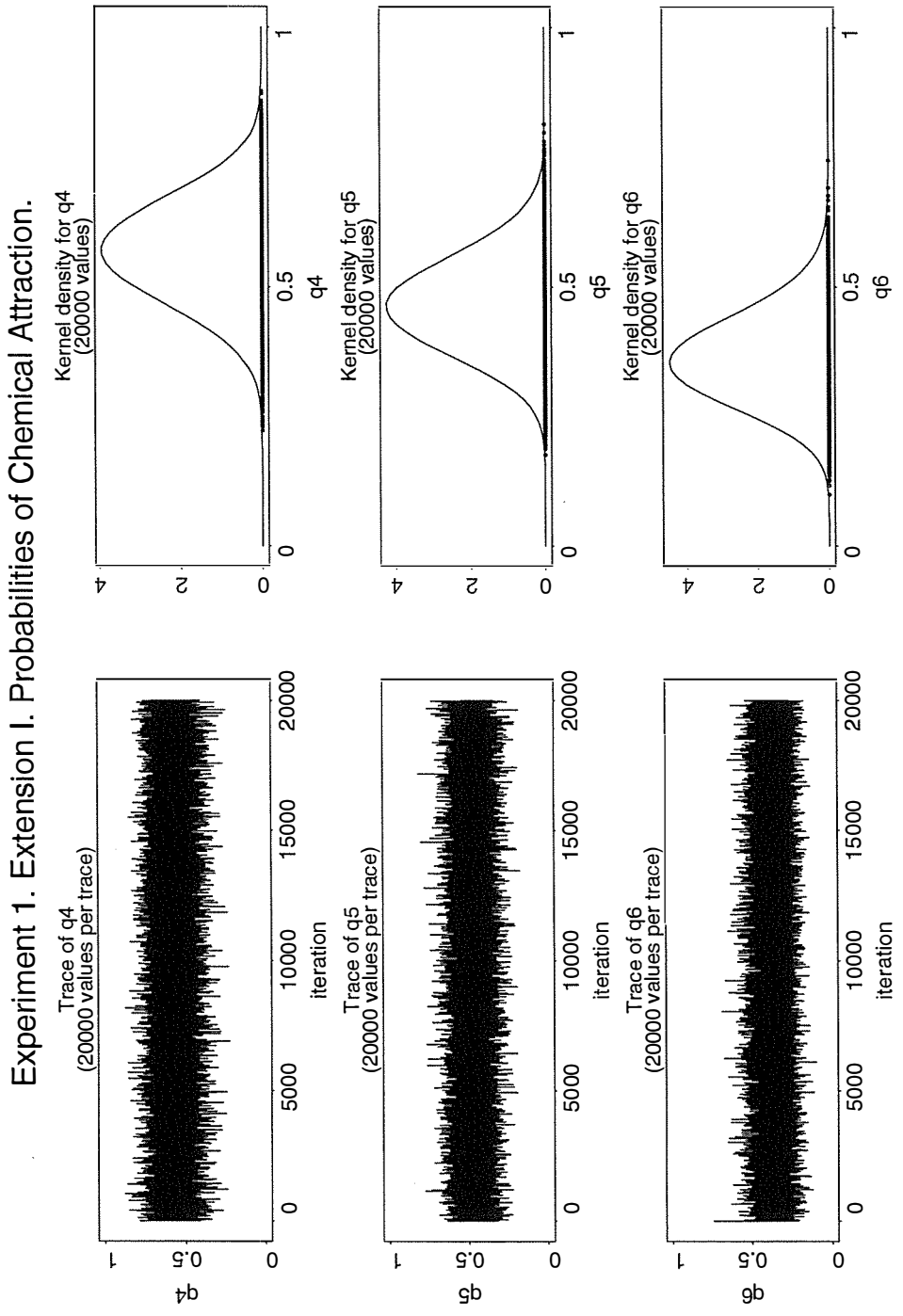


Figure 7.14: Extension I, Experiment 1: MCMC Trace of the posterior distribution of effects of explanatory variables q_4, q_5, q_6.

estimating the mean of each q_k are very similar and of the same order. The integrated autocorrelation times indicate that between 103 and 117 dependent simulations are equivalent to 100 independent simulations, a high degree of efficiency. The 50% credible intervals from the posterior distribution indicate that chemicals 1 and 4 are difficult to distinguish due to almost complete overlap, and that chemicals 5 and 6 can fairly clearly be distinguished from all other chemicals. Finally chemicals 1 and 2 are clearly distinguishable from the other chemicals.

Further diagnostics were investigated to determine whether this convergence was transitory or representative of equilibrium in table 7.4.

Geweke's (1992) convergence diagnostic shows that only q_3 demonstrates a significant difference in the batch means in the first tenth and the second half of the chain. The hypothesis of equivalent distribution of batches in each end of the chain was rejected at the 5% level but accepted at the 1% level for q_3. However, an underlying assumption of Geweke's test is that the posterior distribution it is applied to is normal, and that of q_3 happens to be the least normal of all q_k parameters.

Raftery & Lewis's (1992) test was used to investigate the sample size required to estimate the 2.5th quantile to within ± 0.005 of the true value with probability 95%. The diagnostics reported are N the additional number of iterations needed to achieve this. Approximately 4000 iterations are required to estimate each of the q_k parameters to this level of accuracy. Relaxing the accuracy requirements to estimating the 2.5th quantile to within ± 0.025 or even ± 0.01 of the true value with probability 95% decreased the estimated additional number of iterations required. The dependence factors indicate that the chain was close to independent, with $I \leq 1.17$ for each parameter q_k.

The Heidelberger & Welch (1983) test indicate that stationarity has been achieved for all chains: the Cramer-Von Mises test statistics are all within a manageable range, with the largest value for q_4 corresponding to the greatest deviation from pure Brownian motion; and the Halfwidth test being passed for all k, with the largest value again for q_4 corresponding to the largest halfwidth of the 95% confidence interval of the mean.

Cross-correlations between the q_k parameters are no larger than the negligible value of 0.245 between any pair.

Results transformed to $\text{logit}(q_k) = \alpha_k$ are unchanged in character; all diagnostic test results are identical and descriptive statistics are easier to understand on the q scale rather than the α scale. For this reason results have been reported on the q scale rather than the α scale.

Table 7.1: Extension I, Experiment 1: Descriptive statistics of posterior distribution MCMC simulations of natural statistics for presence/absence $V(x)$. The sample mean and standard deviation are derived from the posterior distribution of the parameters. All of the following diagnostics apply to the MCMC simulated sample obtained from the posterior distribution of the parameter. These are documented and explained in more detail in Section 4.4.7. The SE column contains a triplet of naive, time-series, and batch estimates of SE. Lag 1 autocorrelation of the series and Lag 1 autocorrelation between batches of size 25 are contained in the next two columns. Next is tabulated the Integrated Autocorrelation Time and the 50% credible interval of the posterior distribution.

Parameter	Sample Mean (Stdev)	SE (Naive TS Batch)	Lag 1 AC	Lag 1 Batch AC	IACT	50% Credible Interval
V_0	170 (19)	0.1340 0.2140 0.2280	0.39	-0.010	1.09	[158, 177]
V_1	45.4 (9.49)	0.0671 0.0890 0.0977	0.27	-0.039	1.39	[39,49]
V_2	47.3 (1.17)	0.0825 0.1370 0.1410	0.40	0.065	1.44	[40, 51]

Table 7.2: Extension I, Experiment 1: MCMC Convergence Diagnostics for presence/absence natural statistics $V(x)$. Refer to Section 4.4.7 for more details on diagnostics and interpretation.

		Raftery-	Heidelberger-Welch Tests			
Statistics	Geweke Z	Lewis	CVM* Stationarity		Halfwidth	
$V_0(x)$	-1.59	2,4101	√	0.10	√	0.420
$V_1(x)$	-0.83	2,5211	√	0.02	√	0.174
$V_2(x)$	-0.76	2,5392	√	0.18	√	0.269

* CVM is an abbreviation for Cramer Von Mises statistic

Table 7.3: Extension I, Experiment 1: Descriptive statistics of posterior distribution MCMC simulations of the effects of explanatory variables q_k.

Parameter	Sample Mean (Stdev)	SE (Naive TS Batch)	Lag 1 AC	Lag 1 Batch AC	IACT	50% Credible Interval
q_1	0.56 (0.077)	0.00055 0.00063 0.00066	0.096	0.016	1.17	[0.51, 0.62]
q_2	0.11 (0.044)	0.00031 0.00030 0.00032	0.034	0.019	1.03	[0.08, 0.13]
q_3	0.20 (0.063)	0.00044 0.00047 0.00050	0.056	0.025	1.05	[0.15, 0.23]
q_4	0.57 (0.091)	0.00065 0.00078 0.00082	0.113	0.054	1.16	[0.51, 0.63]
q_5	0.47 (0.083)	0.00059 0.00065 0.00070	0.098	-0.029	1.13	[0.41, 0.52]
q_6	0.36 (0.078)	0.00055 0.00060 0.00063	0.078	-0.003	1.08	[0.31, 0.41]

Table 7.4: Extension I, Experiment 1: MCMC Convergence Diagnostics for effects of explanatory variables q_k

Parameter	Geweke Z	Raftery-Lewis	Heidelberger-Welch Tests Cramer Von Mises Stationarity		Halfwidth
q_1	0.78	3,4380	√	0.13	√ 0.0012
q_2	1.10	2,3853	√	0.10	√ 0.0006
q_3	2.16	2,3919	√	0.15	√ 0.0009
q_4	1.69	3,4350	√	0.31	√ 0.0015
q_5	-0.14	3.4249	√	0.04	√ 0.0013
q_6	1.53	3,4045	√	0.08	√ 0.0012

Table 7.5: Extension I, Experiment 1: Cross correlations between each q_k chain

Variable	q_1	q_2	q_3	q_4	q_5	q_6
q_1	1.000					
q_2	0.127	1.000				
q_3	0.176	0.109	1.000			
q_4	0.245	0.131	0.184	1.000		
q_5	0.234	0.136	0.171	0.244	1.000	
q_6	0.210	0.125	0.166	0.243	0.208	1.000

7.6 Results: EXTENSION I, Experiment 2

The MCMC chains were obtained from a total of 2 million (2,000,000) runs after a burnin of 200,000. A thinning factor of 100 was used, resulting in 20,000 samples used to derive simulations from the posterior distributions. The burnin and thinning factor were selected in order to ensure convergence, according to the diagnostics Geweke's batch statistic, Raftery-Lewis' quantile estimation statistic and the Heidelberg-Welch stationarity tests. Runs of 20,000 were initially considered with zero burnin and no thinning. However all diagnostics were failed by at least one parameter, often several. In particular the autocorrelation estimates and Geweke plots indicated extreme "stickiness" in the chains. Thinning by 100 eliminated most of this problem as well as saving space.

7.6.1 Posterior distribution of θ

As before, we see in Figure D.1 that the posterior distribution is not symmetric in θ components, but rather angled in $\theta_0 + \theta_1$ space, indicating a correlation between prevalence and spatial dependence. High values of temporal dependence were not supported in the posterior distribution.

In experiment 1, prevalence parameter θ_0 value -1.95 was most popular, closely followed by -2.2, with low support for -1.7. In this experiment 2, we chose to investigate the region near -1.95 and -2.2 more closely by selecting fixed values $-1.9, -2.0, \ldots, -2.3$ for investigation. Figure D.1 shows that prevalence parameter value -2.1 was most popular, closely followed by -2.0 and -2.2. This agrees with the results from Experiment 1. The clustering around values of -2.1 suggests that we have located a mode of the posterior distribution. Since the posterior distribution for θ_0 is fairly normal in shape. This indicates that descriptive statistics such as the mean and the standard deviation are good measures of the shape of the posterior distribution of θ_0. The pattern of accepted θ_0 values matches to some extent the pattern of proposed θ^* values. This feature will be explored below in more detail.

Spatial dependence values θ_1 were previously concentrated near 0.5 in Experiment 1. In this experiment we find that values between 0.5 and 1 are well-supported and the posterior distribution is of a narrow fairly Normal shape. Although more extreme values of spatial dependence ($\theta_1 = 0.25$ and 1.25) were proposed quite often, their acceptance rates were relatively low.

Temporal dependence θ_2 between 0.5 and 1.0 were well supported in Experiment 1. Now we find that values of 0.4 are the best supported, with little indication that values of 0.8 or 1.2 are plausible. The proposal distribution of the θ_2 distribution was almost uniform, and the value of 0.4 was obviously better supported by the data.

The marginal distributions are useful in their own right, but do not consider the joint posterior distribution relationships between the θ components. These are examined in the two-way histograms in Figures D.3, D.5 and D.7.

The first bivariate histogram for temporal parameter θ_2 vs spatial parameter θ_1 clearly shows that the marginal pattern of values for the spatial parameter is dominated by the pattern conditioned on temporal parameter value 0.4. Figure D.4 shows the pattern of prevalence parameter values θ_0 associated with each spatio-temporal dependence combination (θ_1, θ_2). For low temporal dependence ($\theta_2 = 0.4$), more extreme levels of prevalence ($\theta_0 \approx -2.3$) correspond to high levels of spatial dependence ($\theta_1 = 1.25$). As spatial dependence decreases ($\theta_1 \to 0.25$), the posterior distribution of prevalence, conditioned on spatio-temporal dependence shifts from being skewed to the right to being skewed the left.

A similar pattern, with a less marked shift in the histogram of prevalence, occurs for the higher level of temporal dependence $\theta_2 = 0.8$. Together these co-plots clearly demonstrate a strong compensatory relationship occurring between the spatial dependence parameter θ_1 and prevalence parameter θ_0.

The bivariate posterior distribution (Figure D.5) of spatial parameter θ_1 and prevalence parameter θ_0 is clearly nearly two-dimensional Gaussian in shape, and matches the Gaussian shaped univariate marginal distributions. There is a slight ridge along a not-quite diagonal traversal of the grid, from (θ_0, θ_1) values of (-2.3, 0.50) towards (-2.1, 0.75). Figure D.6 shows the co-plot of the distribution of temporal dependence θ_2 conditioning on spatial dependence θ_1 and prevalence θ_0. For low prevalence and low spatial dependence (-2.2, 0.25), (-2.3, 0.25), (-2.3, 0.5), the relative distribution of temporal dependence favoured medium values. For all other combinations of prevalence and spatial dependence, the lowest level of temporal dependence was overwhelmingly favoured.

The third bivariate distribution (Figure D.7) for temporal dependence θ_2 compared to prevalence θ_0 shows a similar relationship to that between temporal and spatial dependence as shown in Figure D.3. This is because the marginal posterior distribution for temporal dependence is almost all concentrated on the single value 0.4. In addition, the co-plot for the distribution of the spatial dependence given the temporal and prevalence components (Figure D.8) echoes the pattern shown by the co-plot for the distribution of the prevalence component given the temporal and spatial components shown in Figure D.4.

In Figure D.2 we see that the trace of each θ component appears to be well mixed over time, since the density of simulated values for each value of the θ_l is evenly dispersed along the simulation period.

In Figure D.9 we see that as the overall level of dependence increases, *i.e.* as the sum of spatial and temporal dependence parameters increases, the distribution of prevalence swings from being concentrated at smaller negative values to being concentrated at large negative values. This also provides evidence that we are exploring an interesting part of the parameter space, since this pattern is clearly evident. Thus when spatial and temporal dependence is low, prevalence is high, and vice versa.

7.6.2 Posterior distribution of natural statistics of presence/absence

Corresponding to each of the three components of θ, is one of three components of the natural statistic $V(x)$. The natural statistic estimates are reported for the entire space-time lattice. However a large proportion of sites on this lattice are unknown (corresponding to an inconclusive recorded non-visit $y_{st} = 0$), as tabulated in Equation 7.62.

Descriptive statistics for the whole presence/absence lattice are given in Table D.1. The natural statistic representing the total amount of presence (out of 945 spatio-temporal sites, of which 106 are known presences) is centred at 159 with a relatively tight standard deviation of 13. Natural statistics counting pairs of horizontal and vertical neighbours, V_1 and V_2 respectively, have means of about 44, 38. However the statistic counting vertical neighbours has a standard deviation of 9, and that for horizontal neighbours is 6. This is to be expected due to the larger unknown part of V_1. The lag 1 autocorrelations in the MCMC chains of V_0 and V_1 are very low (both less than 2%) indicating that the chains for presence do not need to be thinned more.

In Figure D.10 the trace plot depicts no consistent changes in the level of the posterior distribution of any of the three components. This evidence is consistent with the conclusion that the MCMC chain is stationary in mean, has converged to equilibrium, and that the

values obtained are therefore representative of the posterior distribution. The posterior distribution of each natural statistic is approximately Gaussian, although in each case, the upper tail is thicker than the lower tail (this feature is obscured by the smooth estimated density curve but is evident from the time series traces.)

Further diagnostics were investigated to determine whether this convergence was transitory or representative of equilibrium. These are documented in table D.2. The Geweke test indicates that there is stationarity in the mean of the simulated marginal posterior distributions of all natural statistics. The Raftery-Lewis estimates of sample size are all below 7,000 which is well within the actual (thinned) 20,000 samples. The Heidelberger-Welch stationarity and half-width tests indicate that the chains are all stationary, and there is sufficient data to estimate the mean accurately.

7.6.3 Posterior distribution of effects of explanatory variables

For each of the unknown α_k parameters in the linear predictor, an MCMC trace of the simulated posterior distribution are shown in Figure D.11 and D.12. Numerical summaries of effects are portrayed on the scale α_k of the linear predictor since these are more Normal than their q_k counterparts. No sustained periods showing non-stationarity in the mean are evident.

Table D.3 gives more descriptive statistics of the posterior simulations. The most powerful probability of chemical attraction (given dingo presence) occurs for close contenders chemical 1 ($q_1 = 0.59, se(q) = 0.067$) and chemical 4 ($q_4 = 0.60, se(q) = 0.081$), with the latter having less precision but slightly higher value. These two chemicals are followed in equal steps by chemicals 5, 6, 3 and 2 (in that order). This coincides with the orderings allocated in virtually all other analyses of this data.

There is sufficient serial correlation to have a small impact on the independent sample estimate of standard error. The naive, time series and batch estimates of standard errors in estimating the mean of each α_k (not tabulated) are very similar and of the same order. The 50% credible intervals from the posterior distribution indicate that chemicals 1 and 4 are difficult to distinguish due to almost complete overlap, and that chemicals 5 and 6 can fairly clearly be distinguished from all other chemicals. Finally chemicals 1 and 2 are clearly distinguishable from the other chemicals.

Further diagnostics were investigated to determine whether this convergence was transitory or representative of equilibrium in table D.4.

Geweke's (1992) convergence diagnostic shows that all q_k chains show similarity of batch means in the first tenth and the second half of the chain. If we examine the distributions of logit (q_k), we find that all posterior distributions are approximately Normal; this result is also supported by QQ plots not shown here.

Raftery & Lewis's (1992) test was used to investigate the sample size required to estimate the 2.5th quantile to within ± 0.005 of the true value with probability 95%. The diagnostics reported are N the additional number of iterations needed to achieve this. Approximately 4000 iterations are required to estimate each of the q_k parameters to this level of accuracy.

The Heidelberger & Welch (1983) test indicate that stationarity has been achieved for all chains: the Cramer-Von Mises test statistics are all within a manageable range, with the largest value for q_4 corresponding to the greatest deviation from pure Brownian motion. The Halfwidth test was passed for all $k \neq 5$. This α_k was the closest to zero (representing $q_k \approx 0.5$, and hence the half-width test would necessarily fail this parameter.

Results reported on scale of $\text{logit}(q_k) = \alpha_k$ ensure that parameters are transformed to

Table 7.6: Extension II, Experiment 1: Design of Θ space.

Model m	θ_0	θ_1
1	-1.70	0.0
2	-1.95	0.0
3	-2.20	0.0
4	-1.70	0.5
5	-1.95	0.5
6	-2.20	0.5
7	-1.70	1.0
8	-1.95	1.0
9	-2.20	1.0
10	-1.70	1.5
11	-1.95	1.5
12	-2.20	1.5

scale best suited to analysis. Descriptive statistics, however, are easier to understand and to compare to previous experiments and models on the q scale rather than the α scale.

7.7 Results: EXTENSION II, Experiment 1

Preliminary results showed that simulation chains of length 2,000,000 were sufficient to ensure convergence of parameter estimates and monitoring statistics for the model of EXTENSION II. A burnin of 200,000 was eliminated from posterior distributions and convergence diagnostics. This was found necessary, from previous analysis, to ensure convergence to the equilibrium. Thinning by 100 was necessary to eliminate excessive autocorrelation in the series and for computing storage requirements.This gave an effective sample size of 20,000 'nearly' independent observations. (See reported values of the IACT which illustrate nearness to independence.) I present convergence diagnostics for chains only of this length to demonstrate that they are suitable for estimating features of posterior distributions of parameters.

An in-depth examination of the application of convergence diagnostics to a similar model was given when presenting results from EXTENSION I. I have therefore relegated figures and tables illustrating these MCMC diagnostics to the appendix, and reference is made to them in the text. Discussion therefore focuses on the outcomes of the MCMC simulations, rather than demonstrating their efficacy and convergence.

Since there were just two components of θ to be examined, this simplified the experiment greatly, and reduced the number of sampled points in Θ space for which the log NC ratios were required. These are a subset of those used with Extension I, but without the extra permutations required for the θ_2 component. They are displayed in Table 7.6 together with the model index m which is referred to in tables and plots.

7.7.1 Posterior distribution of α, β

Convergence diagnostics reviewed in Section 4.4.7 were evaluated for the 6 components of α and for β using a combination of CODA (Best et al. 1995) and SPlus (Becker et al. 1988).

Table 7.7: Extension II, Experiment 1: Summary of Convergence diagnostics

MCMC Diagnostic		Comments	Appendix	
Type	Name	on Convergence	Figure	Table
Run	Raftery-Lewis	Run length required for all α_k and β parameters is less than 4000.		E.2
Length	IACT	Indicates that with thinning, within each chain, simulations are effectively almost independent.		E.1
	Trace plots	Simulated time series appears stationary and convergent. Posterior distributions all appear very Gaussian (supported by QQ plots).	E.1,E.2	
Stickiness,	SEs	Not much difference between time-series and naive estimates indicates near independence.		E.1
Thinning	ACF	Autocorrelation for orders other than 0 are not significant.	E.3	E.1
	CUSUM	No evidence of non-stationarity or lack of hairiness.	E.4	
Stationarity,	Geweke	Consistent with hypothesis of stationarity across batches.		E.2
Mixing	HW tests	Evidence of stationary distribution and only α_5 showed evidence of excessive variability.		E.2

Results are tabulated in Table 7.7 to indicate whether the chain for each parameter passed each diagnostic test. Reference is made in this table to the figures and other tables located in the appendix which support these results.

Here we discuss the 7 parameters $\alpha_1, \alpha_2, \ldots, \alpha_6, \beta$ which enter the linear predictor for $q_{ist} = \mathrm{E}[\, y_{ist} \mid z, \alpha, \beta \,]$ via

$$\mathrm{logit}\,(q_{ist}) = \begin{cases} \alpha_{\tau_{ist}} + \beta y_{is,t-1}, & z_{s,t-1} = 1, \quad z_{st} = 1 \\ \alpha_{\tau_{ist}}, & z_{s,t-1} = 0, \quad z_{st} = 1 \\ 0, & z_{s,t-1} = 0, \quad z_{st} = 0 \end{cases}$$

We can rewrite this in tabular form as Table 7.8. We use card suits to denote each of the four situations, having run out of roman and greek letters and numbers! The four situations may be described as follows. Situations ♣ and ♠ are the only situations with positive probability of a dingo visiting, since a dingo must necessarily be present first. Situation ♣ represents the case where, at a particular site-time combination, a dingo is present today

Table 7.8: Extension II: Categories of situations affecting the distribution of y_{ist} given y, z.

Situation	$y_{is,t-1}$	z_{st}	$z_{s,t-1}$	$\eta_{ist} = \text{logit}(q_{ist})$
♣	1	1	1	$\alpha_{\tau_{ist}} + \beta$
♠	0	1	*	$\alpha_{\tau_{ist}}$
♡	0	0	*	0
♢	1	0	1	0

Table 7.9: Summary of posterior distribution for expected probability of a visit given that a dingo is present, and depending on whether there was a previous visit and/or presence.

	50% credible interval and median for q_{ist}					
Chemical	$\text{E}[y_{ist} = 1 \mid ♣]$			$\text{E}[y_{ist} = 1 \mid ♠]$		
τ_{ist}	25%	50%	75%	25%	50%	75%
A	0.55	0.60	0.66	0.43	0.50	0.58
B	0.07	0.10	0.13	0.04	0.06	0.09
C	0.15	0.19	0.23	0.09	0.13	0.18
D	0.54	0.61	0.68	0.43	0.51	0.59
E	0.43	0.50	0.56	0.32	0.39	0.47
F	0.32	0.37	0.43	0.22	0.28	0.35

and yesterday, and a dingo visited yesterday. Situation ♠ represents the case where a dingo is present today, but no dingoes visited yesterday (and a dingo may or may not have been present yesterday). Situation ♡ is the exact opposite of situation ♣ and represents the case where dingoes are absent today, no dingoes visited yesterday (and dingoes may or may not have been present yesterday). Situation ♢ is the same as for situation ♣ except that dingoes are absent today (*i.e.* a dingo visited and was present yesterday).

The summary statistics computed on the scale of the linear predictor are provided in Appendix Table E.2. These can be transformed to credibility intervals on the scale of the responses due to the one-one monotonic increasing relationship between q_k and α_k. On this scale, the intervals can be compared to results from the Frequentist analyses in Chapter 3 and to the Pilot experiment in Chapter 5. Summary statistics for the posterior distributions of q_k matching the situations above are presented in Table 7.9. Given situation ♣ then the probability of a visit is between 10 and 50% higher than for situation ♠. Chemicals A and D are 60 and 50% likely to attract a dingo given a dingo is present and given previous visit.

For the ♣ case where $y_{is,t-1} = z_{st} = z_{s,t-1} = 1$, the linear predictor involves more than one parameter, *i.e.* $\eta_{ist} = \alpha_{\tau_{ist}} + \beta$. A credible interval can be computed for this linear predictor by compiling of the joint posterior distribution of α_k and β, from this constructing the distribution of $\alpha_k + \beta$ and then using the sample 0.025 and 0.975 quantiles to construct the credibility interval for this quantity. This contrasts with a method which can proceed if the joint posterior distribution of α and β is not available. A simplified method is obtained by noting that the posterior densities of both the α components and β are approximately Normal. Then an approximate 95% credible interval can be constructed via $\eta_{ist} \pm 1.96$ s.d.(η_{ist}) where

$$\text{s.d.}(\eta_{ist} \mid ♣) = \text{s.d.}(\alpha_k + \beta) = \sqrt{\text{Var}[\alpha_k] + \text{Var}[\beta] + 2\text{Cov}[\alpha_k, \beta]}. \quad (7.63)$$

where $k = \tau_{ist}$. Sample estimates obtained from MCMC simulations can replace the population quantities. Since all the covariances between β and each of the α components are in fact negative though not large (Table E.3), then this means that the accuracy of the linear predictor in this case is in fact improved. However, it does indicate that the joint distribution should provide better information about these sums, since the summands are not independent or else these covariances would have been very small.

The posterior distribution of β is itself of interest, since it has a negative-valued mean -0.421 with standard deviation 0.415. Thus the conditional probability of a visit given a visit in a previous time period is *lower* than the conditional probability of a visit given no visit in the previous time period. This supports exploratory data analysis presented in Chapter 3 which noted that there were higher number of visits in alternating days. Thus there is a carryover effect from one day to another, in that dingoes are less likely to visit a site, generally for all chemicals, if it was visited the previous day.

7.7.2 Posterior distribution of dingo presence

Presence/absence must be modelled at the level of individual sites in order to gauge the effects of explanatory variables. For investigation of MCMC convergence diagnostics, however, it is simply not feasible nor informative to assess nearly 1000 time series of the presence and absence process z. The $V(z)$ summary statistics of this process are the canonical statistics which define the prior distribution, and so are excellent candidates both for monitoring convergence as well as for summarizing the posterior distribution. Note that for every recorded visit, there was necessarily a presence, and no modelling was required; inference for the unknown $\widetilde{z}_{st}$ was only required where no visits were recorded at a site.

Results from applying MCMC diagnostics are summarized in Table E.7 and Figures E.6 and E.5. Run length was adequate (indicated by IACT, trace plots, and Raftery Lewis) for achieving equilibrium for the posterior distribution of presence/absence statistics, and for estimating the 0.025 and 0.975 quantiles accurately. Thinning was sufficient since naive and time-series estimates of standard errors were very similar. In addition, Lag 1 autocorrelation was computed to be 0.10 for both V_0 and V_1, with other lag autocorrelations being below 0.01. This indicates that the chain was not too sticky. CUSUM plots and the hairiness diagnostic, together with Geweke's test statistic and the Hiedelberger-Welsh tests all indicated that the distribution was stationary and that the chain was well mixed.

Figure E.7 shows the probability of changing the V components averaged cumulatively over iterations. This is analogous to the acceptance probability shown for parameters updated using Metropolis-Hastings samplers. Here each $\widetilde{z}_{st}$ component is updated using a Gibbs sampler, with only 2 possible outcomes due to its binary nature: retain the current value or 'flip'. Thus the aggregated $V(z)$ statistic changes its value from one iteration to the next depending on whether the site updated at that iteration was flipped or not. The prevalence statistic $V_0(z)$ changes just under 40% of the time, and the spatial clustering statistic $V_1(z)$ changes just under 50% of the time. Inspection of trace plots of a number of individual z_{st} parameters showed that these flips were occurring at sites well dispersed along the transect.

Overall 106 out of the 945 site-time combinations were visited during the experiment as shown in equation (7.62). These corresponded to definite dingo presences in the underlying spatio-temporal process z. Therefore, of the median number of presences (162 in Table E.6), approximately two-thirds of these resulted from observed visits, the other third were imputed by the model. Since the potential number of presences is 945, this represents quite

a significant increase. Further investigation of the dingo data reveals that visited adjacent pairs of sites numbered 21. Therefore, of the median number of presences to adjacent sites (45 in Table E.6), approximately one-half resulted from observed visits, the other half were imputed by the model. Since the potential number of site-paired presences is 938, this represents quite a significant increase.

The posterior distribution of the V_0 and V_1 statistics appeared very Gaussian. The expected values of these statistics were estimated with a great deal of accuracy from these simulations (with standard errors of less than 1%.)

7.7.3 Posterior distribution of θ

The probability that a particular $\theta_{(m)}$ was proposed obviously depended on the probability that its neighbours were accepted, as shown in Figure E.8. Thus although models 4 and 5 were accepted often, they were not often proposed, and similarly for models 8 and 9. Models 6 and 7 were rarely accepted, and were not often proposed either. Models that were proposed the most often were models 1,3 and 12 yet were among those models most often not accepted. Figure E.9 shows that, like model extension I, there is a tradeoff occurring between prevalence and spatial dependence. Models with a consistent sum of $\theta_0 + \theta_1$ are most frequent in the posterior distribution.

Posterior distribution statistics are summarized in Table 7.11.

A time series plot of model indices m is a one-dimensional way of inspecting the traversal of the chain through two-dimensional θ space. To enhance the visual accessibility of the plot, MCMC iterations were further thinned by 50 to give a trace through m space in Figure E.10. Although only 5 models were clearly preferred there was a good level of mixing throughout the simulation, for both accepted and proposed values of m, and therefore of θ.

The usual diagnostics were not found to be useful for an enumerative variable such as m which represents a small discretized two-dimensional space. An inspection of 'moving' transition probabilities or 'moving' histograms computed over the time period, similar to the way that moving averages are computed, might be useful for a more complex situation than this. In addition, multivariate techniques for reducing dimensionality of data, such as principal components, procrustes rotation or correspondence analysis, might be useful for more complicated multi-dimensional data.

Table 7.10: Descriptive statistics for the simulated posterior distribution of each component of $V(x)$

Statistic	Q_1	Average	SD	Median	Q_3	95% *c. i.* for mean
$V_0(z) = \sum_{st} z_{st}$	151	163	(16.4)	162	175	[162.76, 163.24]
$V_1(z) = \sum_{st} z_{st} z_{s+1,t}$	30	46	(10.8)	45	53	[45.94, 46.26]

Table 7.11: Descriptive statistics for the simulated posterior distribution of θ

	Two-dimensional θ		Frequency	
Model m	θ_0	θ_1	accepted	proposed
1	-1.75	0	0.05	0.11
2	-1.75	0.5	0	0.08
3	-1.75	1	0	0.13
4	-1.75	1.5	0.27	0.07
5	-1.9	0	0.25	0.04
6	-1.9	0.5	0	0.05
7	-1.9	1	0.01	0.04
8	-1.9	1.5	0.24	0.06
9	-2.2	0	0.17	0.1
10	-2.2	0.5	0	0.11
11	-2.2	1	0	0.07
12	-2.2	1.5	0.01	0.13

7.8 Discussion

In this section I have considered two extensions to the hierarchical model introduced in Chapter 5, which modelled success/failure data as conditionally independent given an underlying model for spatially and temporally dependent presence and absence. A specific binary Markov random field model, the autologistic distribution, has been examined in detail as a candidate for this underlying spatio-temporal dependence model. The three-tier model with these features was investigated in Chapter 5. A major limitation was that it conditioned on parameters governing the underlying spatio-temporal model for presence.

This limitation, essentially a problem of model choice, can be overcome by extending the three-tier model to a four-tier model. The additional tier models the distribution of these spatio-temporal dependence parameters. In a Bayesian context, this amounts to using a hyper-prior distribution to describe these parameters which enter into the prior distribution for spatio-temporal dependence.

Two different extensions to the three-tier model were considered. Each treated temporal dependence differently. The first extension retained the approach of the basic three-tier hierarchical model, where temporal dependence affects presence and absence, thereby indirectly influencing success and failure. The second extension instead considered that temporal dependence would affect successes and failures directly. Here I discuss and compare the relative merits of each extended model, Extension I and Extension II.

- Estimates for the probability of attraction to k given presence of dingos were similar for both extended models. This indicates that it is not so important where temporal and spatial dependence is accounted for in the model, although it is important to model both temporal and spatial dependence.

- Note the significant increase in computational complexity encountered when modelling temporal dependence in the underlying spatio-temporal model via an autologistic distribution, compared to modelling temporal dependence within the linear predictor for the success/failure data model. Balancing this with the similarity in results using either method of modelling temporal dependence, I conclude that the second method Extension II is preferred.

- Favoured chemicals were 1 and 4, having attraction probabilities of 0.56 and 0.57 under EXTENSION I, or 0.55 and 0.55 under EXTENSION II. These were followed at equal intervals by chemicals 5, 6, 3 and 2, in order of decreasing attraction probabilities. Estimates of precision were similar under both model extensions, although EXTENSION II tended to produce less correlated MCMC simulations, as evidenced by the integrated autocorrelation time (IACT) and raw and batch lag 1 autocorrelations.

- In EXTENSION I, non-zero spatial and temporal presence dependence parameters θ_1 and θ_2 were well supported by the data. Similarly for EXTENSION II, estimates of θ_1 and β for spatial dependence in presence and temporal dependence in attraction were also different from zero. Interestingly, the time dependence parameter obtained in the Extension II model was negative, indicating a negative carryover effect on visits from one day to the next. In the Extension I model, there was a small positive carryover effect on presences for consecutive days. Negative carryover effects over time were not investigated for the Extension I model. This leads to a subtle though important difference in interpretation, requiring clear distinction between visits and presence.

- As normalization constants had to be estimated off-line using extra MCMC simulations, only a small discretized portion of θ space could be investigated. In EXTENSION II, prevalence closer to -1.95 than -1.7, and spatial dependence between 0.5 and 1.0 was supported by the data. In EXTENSION I, the posterior modes of θ were $(-1.95, 0.5, 0.5)$ or $(-2.2, 0.5, 1.0)$ for prevalence, space and time components. The joint posterior shows a trade-off or negative correlation between space and time dependence.

- This negative correlation between spatio(-temporal) dependence and prevalence indicates that perhaps another parameterization should be considered, such as the Ising parameterization. This model focuses on similarity/dissimilarity between neighbours of a given type rather than the number of presences in pairs of a given type. It therefore focuses less on explaining presences, and more on explaining whether absences are grouped with absences, and presences with presences.

- The natural statistics V summarize the dingo presence parameters. Little difference was found when the first two components common to both extended models were compared between extended models. This indicates similar overall presence patterns and pairs of presences within spatial neighbours were obtained under each modelling extension.

- All parameters in the models passed the various *ad hoc* MCMC convergence diagnostics: visual trace; Geweke's batch-stationarity test assuming normal-like posteriors, Raftery & Lewis's goodness-of-fit to quantiles convergence diagnostic; and the suite of time-series based stationarity tests of Hiedelberger and Welch.

- Some innovative approaches to visualization allowed inspection of posterior distributions of the 2- or 3- dimensional discretized θ component.

The work of Chapters 4 through to the present one have shown that it is possible to proceed beyond the first model investigated in Chapter 3, where underlying spatio-temporal presence was explained by a set of independent variables representing the probability of presence.

The model extensions of Chapter 7 ensure that the basic hierarchical model of Chapter 5 could be augmented to incorporate inference for the underlying presence process. This has demonstrated that a full hierarchical version of the model introduced in Chapter 5 is indeed feasible.

Chapter 8

Conclusions

Contents

8.1 Summary of methodological results 274
8.2 Comparison of results for *Dingo* case study 275
8.3 Future directions 278

8.1 Summary of methodological results

> 'Begin at the beginning', the King said, gravely, 'and go on till you come to the end: then stop.'
> *-Lewis Carroll Alice's Adventures in Wonderland (1865)*

The *dingo* case study provided a useful focus for the methodological issues addressed by this thesis. The dataset was of manageable size, yet research questions were sufficiently challenging to raise statistical modelling issues of varying complexity. The problem of ambiguous zeroes in presence/absence data is prevalent in many situations, including the *dingo* case study and many biogeographical applications which involve mapping of flora or fauna. Chapter 2 described these application areas, particularly those involving computer generated imagery, which have similar features to the *dingo* case study. This demonstrates that the methodology presented in this thesis can potentially be applied to a technology which is becoming increasingly valuable in today's information-driven society.

The Frequentist approach of Chapter 3 tailored existing and standard statistical methodology and computational techniques. In order to achieve this, however, the underlying spatial and temporal presence process was based on independent variables. Ad-hoc methods provided more accurate estimates of standard errors. These limitations provided an impetus for investigation of models which captured the dependent nature of the spatial and temporal presence process. This initial chapter provided a benchmark and motivation for further work to improve the modelling and develop supporting methods for inference.

Binary Markov random field models were identified as suitable for modelling a dependent spatio-temporal presence process in Chapter 4. Of particular interest is the autologistic distribution, becoming more visible in its single parameter form in the spatial statistics and image analysis literature. I found the three-parameter autologistic distribution to be more suitable for spatial statistics applications such as those outlined in Chapter 2. With this parameterization there are two extra parameters. A prevalence parameter allows presences and absences to occur on the lattice with unequal marginal probability. This parameterization also allows differentiation between dependence in two directions, which in the *dingo* case study amounted to temporal dependence and spatial dependence along a one-dimensional transect. In situations involving a two-dimensional spatial lattice, the North-South and East-West spatial dependence can be examined separately, for example. In biogeographical situations such as mapping flora and fauna, for example, it is very common for meteorological and geographical conditions to vary greatly in these two directions. For example, in South-East Queensland of Australia, mountains align along the coast, running approximately North to South. This greatly affects meteorological conditions such as rainfall and humidity, and wind conditions.

A Bayesian approach to inference of binary MRFs, based on MCMC computational techniques, was found to be effective. Hence the Bayesian approach was considered for application to all aspects of the hierarchical model. This hierarchy consisted of three tiers, for conditionally independent success/failure observations, given an underlying presence process and covariates, based on known spatio-temporal dependence parameters. Preliminary investigations were made into the use of a Bayesian approach to analysis of this three-tier model in Chapter 5. At this stage, model choice was hampered due to the fixed nature of spatio-temporal dependence parameters. The addition of a fourth tier to the hierarchy was posited at the end of this chapter. This posed a rather challenging problem of estimating normalization constant ratios, to which the entirety of Chapter 6 was addressed.

A literature review of methods for estimating NC ratios of binary MRFs, at the beginning of Chapter 6, showed that in many cases the problem had variously been avoided, deemed infeasible or addressed using methods which compromised the underlying spatio-temporal dependence structure. A relatively simple idea was developed which had been given some consideration in the statistical physics literature, and more recently in parallel work on path sampling methods in the statistical literature (Gelman & Meng 1998). This method, which I have called the Integrated Mean Canonical Statistic method, is based on an estimating equation which equates theoretical and empirical expressions for the mean canonical statistic of binary MRFs. Again MCMC computational techniques were found to be useful for estimating empirical values of the mean canonical statistic. Results were compared to the more computationally demanding Reverse Logistic Regression method of Geyer (1996), as well as simpler methods based on Monte Carlo simulations.

Given the ability to estimate normalizing constant ratios, the extended models of Chapter 7 ensured that the basic hierarchical model of Chapter 5 could be augmented to incorporate inference for the underlying presence process. They illustrated that the hierarchical model was flexible enough to permit temporal dependence to directly impact on either observed success and failure or the latent presence and absence process.

8.2 Comparison of results for *Dingo* case study

We may compare results for the *Dingo* case study, as obtained from all methods considered in this thesis: extreme conditional methods (Chapter 3), the Frequentist approaches based on either blocked or smoothed estimates of the probability of dingo presence (Chapter 3), the basic Bayesian approach using a three-tier model (Chapter 5) and the two extensions using a four-tier model (Chapter 7). The estimates (and standard deviations) of chemical attractiveness parameters for Frequentist models and the corresponding descriptive statistics of the posterior distributions of these parameters for Bayesian models are summarized in Table 8.1. Note the subtle differences in interpretation of these estimates. In the Frequentist context, these represent the range of parameter estimates expected over many random replications of the data. In the Bayesian context, these represent the range of (random) parameter values supported by the dataset observed.

The extreme unconditional and conditional models are denoted models EU and EC respectively. The Frequentist models based on an underlying presence process comprising independent variables use blocked (model FB) and smoothed estimated probability of presence (model FS). The latter model has standard errors which are bootstrapped using an estimated marginal distribution for Day 1 (model FS1) or using observed data for Day 1 (model FSC). The Bayesian models based on an underlying presence model with spatio-temporal dependence are prefixed B and denoted according to the Extension number (I or II) and according to the experiment (1 or 2).

From these results it is obvious that the inclusion of a dependent spatio-temporal process has significantly altered the estimates of success of chemical lures. The main features changed can be summarized as follows:

- the highest-ranked chemical(s)
- the success of the highest-ranked chemical(s)
- the middle-ranked chemical(s)

Table 8.1: Summary of chemical attractiveness, over extreme conditional and unconditional models and Frequentist models of Chapter 3 and the Bayesian models of Chapter 7.

	Method	$q_k = \mathrm{E}[y_{ist} = 1 \mid z_{ist} = 1, \tau_{ist} = k]$						Comments
		q_A	q_B	q_C	q_D	q_E	q_F	Comments
		$\mathrm{sd}(q_A)$	$\mathrm{sd}(q_B)$	$\mathrm{sd}(q_C)$	$\mathrm{sd}(q_D)$	$\mathrm{sd}(q_E)$	$\mathrm{sd}(q_F)$	
EU	Unconditional: presence always assumed	0.121 (0.018)	0.016 (0.007)	0.029 (0.009)	0.098 (0.017)	0.083 (0.016)	0.060 (0.013)	$z_{st} = 1 \; \forall s, t$
EC	Conditional: no presence assumed when no visits observed	0.478 (0.089)	0.0851 (0.041)	0.184 (0.070)	0.538 (0.109)	0.406 (0.095)	0.388 (0.093)	$z_{st} = \max_i y_{ist}$
FB	Block presence ($n_b = 9$ sites per block)	0.555 (0.164)	0.076 (0.044)	0.151 (0.089)	0.497 (0.215)	0.413 (0.179)	0.323 (0.161)	$z_{st} = 1$ w.p. p_{bt} $(b-1)n_b + 1 \leq s \leq bn_b$
FS1	Smoothed presence (over 9 sites) with estimated first day	0.510 (0.082)	0.074 (0.033)	0.148 (0.053)	0.472 (0.097)	0.381 (0.085)	0.301 (0.074)	$z_{st} = 1$ w.p.p_{st} $p_{st} = ...$
FSC	Smoothed presence (over 9 sites) conditioning on first day	0.532 (0.068)	0.075 (0.030)	0.152 (0.051)	0.484 (0.081)	0.391 (0.070)	0.313 (0.064)	$z_{st} = 1$ w.p.p_{st} $p_{st} = ...$
BI1	Extension I (time enters $p(z\|\theta)$ as θ_2)	0.56 (0.077)	0.11 (0.044)	0.20 (0.063)	0.57 (0.091)	0.47 (0.083)	0.36 (0.078)	$p(z\|\theta) \sim \mathrm{AL}(\theta_0, \theta_1, \theta_2)$ $\eta_{ist} = \alpha X$
BI2	Extension I (Experiment 2)	0.59 (0.067)	0.11 (0.064)	0.20 (0.080)	0.60 (0.081)	0.49 (0.080)	0.38 (0.085)	
BII1	Extension II (time enters $\mathrm{E}[y_{ist}\|y_{is,t-1}, \ldots]$ as β)	0.55 (0.075)	0.10 (0.041)	0.19 (0.058)	0.55 (0.089)	0.46 (0.079)	0.35 (0.074)	$p(z\|\theta) \sim \mathrm{AL}(\theta_0, \theta_1)$ $\eta_{ist} = \alpha X + \beta y_{is,t-1} z_{is,t-1}$, $z_{st} = 1$

- the success of the lowest-ranked chemical
- estimates of standard deviations for the estimates $\widehat{q}_k$
- relative estimates of standard deviations for the estimates $\widehat{q}_k$

With the extreme conditional and unconditional models, as well as the Frequentist models, Chemical A is clearly ranked the most successful, with Chemical D also clearly ranked second. In contrast, the Bayesian models rank Chemical D as slightly more or equally successful compared to Chemical A. Thus accounting for underlying spatio-temporal dependence dampens the success of Chemical A but promotes Chemical D's success. The number of observed failures to visit Chemical D can therefore be more attributed to absence of dingoes in the vicinity rather than a non-response caused by failure of the chemical to attract dingoes. On the other hand, the number of observed failures to visit Chemical A are less likely to be due to absence than non-response (in comparison to Chemical D.)

With the extreme conditional model, the second and third ranked chemical lures are Chemical D and E respectively, and are very close in their ability to attract. All other models show a clear distinction between the top-ranked chemicals A and D in comparison to Chemical E. Thus if we take the extreme conservative approach, and only consider those sites which we know for certain were visited, it is difficult to differentiate between Chemicals D and E. In all other models where more information is taken into account these chemicals can be clearly differentiated.

The fourth ranked chemical F is approximately twice as successful as fifth ranked chemical C, which is again about twice as successful as the sixth ranked chemical B. This statement most closely matches the situation for the Frequentist models and for the extreme unconditional model. The extreme conditional model has slightly wider gaps between these three lowest ranked chemicals, and the Bayesian models have slightly lesser gaps. Hence taking into account spatio-temporal dependence in presence makes it more difficult to differentiate between the three least successful chemicals, implying that for least successful chemicals C and B, absence is responsible for more failures than non-response when compared to chemical F.

The extreme unconditional model overestimates the number of replications over which probabilities of success can be estimated, and does not take into account any dependence between observations, leading to the smallest apparent estimates of standard deviations in these probabilities.

In general, estimates of standard deviations of the posterior marginal distributions of q_k in the Bayesian models are more consistent over chemical index k, in comparison to the way in which standard deviations vary with the magnitude of $\widehat{q}_k$ for Frequentist models or in the extreme models. This could be attributed partly to the separation of errors which occurs with the Bayesian hierarchical models.

Within the pair of Bayesian models based on the same Θ space (Experiment 1), BII has slightly smaller standard deviations and attractiveness estimates than BI1. Modelling temporal dependence as impacting on observed visits imparts more information than when it is modelled as impacting on the unobserved presence process.

Experiment 2 focuses on a smaller Θ space than Experiment 1 for Bayesian model Extension I. For BII, the standard deviations of attractiveness of the most successful chemicals (A and D) are lowered and those of the least successful chemicals (B and C) increased, when compared to BI1. The estimates of chemical success are slightly higher for Experiment 2 than those with the more extensive Θ space Experiment 1. Thus when the spatio-temporal

dependence space is restricted to those values considered most likely, estimated chemical attractiveness is increased, and their standard deviations are more evenly balanced between chemicals.

We may conclude in the *dingo* case study that a three-tier model can give limited preliminary results on the relative success of each chemical lure. These lures retained a roughly similar pattern of success across a range of models allowing for various levels of dingo prevalence, and for different spatio-temporal patterns of dingo presence. However in order to determine the extent of this success, necessary for any cost-benefit analysis of the implementation of these chemical lures, a four-tier model was required. This showed that a dingo prevalence parameter representing a level somewhat in the middle of *a priori* expectations was best supported by the data, together with medium levels of spatial dependence and slightly higher levels of temporal dependence in dingo presence.

8.3 Future directions

Inference was found to be computationally intensive due to the need to estimate normalization constant ratios offline. The work of Denham & Mengersen (1999), based on satellite imagery comprising millions of pixels, presents new challenges for large datasets. It highlights the need to refine computational efficiency of the algorithms presented here. Simulation methods, such as the block-update method, umbrella sampling method and Wolff's algorithm mentioned in Chapter 4 show potential for achieving this increase in efficiency. Furthermore, a more hierarchical approach to MCMC design may be appropriate, along the lines of Reversible Jump algorithms (Green 1995, Carlin & Chib 1995) or bridge sampling (Valleau & Card 1972, Gelman & Meng 1998). This may be more efficient for jumping between θ_m values.

In addition it may be possible to tailor some of the more recent MCMC diagnostic methods (such as those reviewed in Mengersen et al. (1999)) to binary MRFs to provide more precise estimates of time to convergence. In particular the perfect sampling approach has been applied to MRFs and could provide a useful avenue for improving efficiency.

Another feature prevalent with satellite imagery data is that of ordinal rather than binary observations. Markov random fields, such as the Potts model, have been used in image analysis and provide good candidates for an extension of the methods presented here to ordinal data. Many measurement error models concerned with deriving true images based on noisy images have been studied in the image analysis field.

Different formulations for the impact of covariates on the responses may be considered. In this thesis a linear predictor with logit link function has been considered in detail for the *dingo* case study, although the framework provided is suitable for the gamut of link functions available for generalized linear models (McCullagh & Nelder 1993). A probit link function with underlying autoGaussian model has already been considered in Weir & Pettitt (1999). Generalized additive models are a relatively new innovation that may be applied to the predictor involving covariates.

In some contexts, including satellite imagery data, the zero-inflated data problem could be more complex than that considered here. Zeroes can arise from a variety of missing value or threshold processes; here we have considered only one source of missing values, that of absence, and one source of non-response, *i.e.* non-attractiveness of chemicals.

The supervisor has suggested, in the case of an extremely large finite lattice, that spatio-temporal dependence parameters might be estimated from smaller sub-lattices and then

somehow aggregated to represent parameters of the super-lattice. This then raises the question of how parameters of finite lattices are related to those with similar spatio-temporal dependence patterns but different lattice size. Many theoretical results on the properties of binary MRF models such as the Ising model focus on the infinite lattice case in statistical physics contexts. However there are some results available for finite lattices, particularly small lattices. This would require further literature review and theoretical investigation.

A further extension would enhance the representation of multivariate relationships. For instance, multivariate responses observed at each space-time position can be modelled as dependent, with some variance-covariance structure, instead of as independent.

A feature of the work of Chapters 6 and 7 was that the parameter space for the autologistic model was discretised to give a small set of distinct values. This permitted off-line computation of log normalization constant ratios. The investigation proceeded in stages considering differently defined discretisations as more information became available about the posterior distribution. Using the methods of Gelman & Meng (1998) it should be possible to formulate a Markov chain so that both the posterior distribution and the sampling distribution of the data is the target distribution for the chain, allowing the parameter space to be considered as continuous and investigated simultaneously as the normalization constant is estimated.

We have demonstrated the feasibility of modelling a spatio-temporal component of a model using a three parameter autologistic model. Although the data layer of the model was specialised to binary data, there is no practical restriction to substituting the binary data model by a more complex data model involving count data, continuous data, categorical data, multivariate data, and so on. As such the research has broken out of the mould where the autologistic distribution is used with approximate inference methods, due to the difficulties imposed by the need to calculate the normalization constant, in a computationally complex fashion. This is the first time a fully Bayesian approach to inference has been implemented, in the literature as far as I am aware, for modelling binary data observed on a two-dimensional lattice.

Extension II of Chapter 7 is less complex than Extension I of Chapter 7 because the time dependence is modelled through the data model, rather than through the underlying spatio-temporal process model. Very similar results are obtained from Extension I and Extension II for the effects common to both models. This is reassuring. It also raises the possibility that a two-dimensional space-time model could be constructed using the three-parameter autologistic model for the underlying process describing spatial dependence and a time dependent observation driven model with no more computational complexity than involved in Extension I.

Appendix A

Graphical representation of models

Modelling options can be represented graphically using a graphical method proposed by Whittaker (1990). These include Directed Acyclic Graphs (DAGS) and Undirected Local Markov Graphs; both types of graphical representations of models will be used for models applied in this thesis.

These graphs comprise different shaped nodes connected by different types of lines or vectors, grouped according to plates. Different types of modelling objects are represented by different shapes of nodes:

Object	**Node**	Other features
Constant	Rectangle	no parents
Stochastic (data or parameter)	Circle	assigned a probability distribution; target of solid arrows
Logical (Deterministic)	Circle	logical function of other node; target of dashed arrows
Frame	Dashed rectangle	groups together objects with the same error structure, which usually have the same subscripts

The direction of arrows linking objects indicates the direction of dependence. If X depends on Y then the arrow joining the two nodes will point at Y. A line is used to join two objects when dependence is undirected, which occurs for example when they have a joint distribution function (*e.g.* Markov Random Fields). However one can always write $p(A, B) = p(A|B)p(B)$ and change the relationships within the model.

A plate then encompasses objects having the same error structure, usually objects having the same subscript.

Appendix B

Frequentist analysis of *Dingo* case study: Bootstrapped estimates of standard errors

B.1 Bootstrap estimates of sample errors

The exact standard errors for the EM analysis cannot be obtained since the exact distribution of the estimates is unknown. Instead we can approximate standard errors via the asymptotic information matrix or via bootstrapping the original data.

An asymptotic estimate of the variance-covariance matrix can be based on an approximation of the observed information matrix using the complete data information matrix and the variance-covariance matrix of the complete data. We give details in this appendix on how to obtain standard errors for the estimates using this method.

An alternative is to generate samples 'similar' to the original data via bootstrapping (Efron 1979). Parameter estimates obtained from each individual bootstrap sample can be combined to approximate the distribution of these estimates. Many statistics, including standard errors and quantiles can be obtained from this approximate distribution.

We need to determine exactly how bootstrap samples are to be generated from the original data. A simple method would be to use the design matrix to categorize site-time-location triplets according to chemical treatment, and then draw bootstrap samples stratified by chemical category. This ignores the 'repeated measures' aspect of the design whereby chemicals were repeatedly assigned to site-location tuples for each time point.

We have opted for a more sophisticated approach and generate bootstrap samples

$$y^* = \{y^*_{ijt}; \quad i = 1, 2, \ldots, 135; \quad j = 1, 2; \quad t = 1, 7\}$$

for dingo visits within chemical pair categories according to a simple Markov Chain model. There were three strategies for generating the bootstrap sample for the first day's data, as discussed in Sections B.1.2 and B.2. The bootstrap sample for remaining days was generated from a simple one-step Markov Chain model for visits conditioned on the previous day's visit.

B.1.1 Bootstrap Sampling Methodology

Consider separately each group of sites having a particular chemical pair, that is sites $\{i : \tau_i = (A, B)$, where $\tau_i = (A, B)$ denotes that $\tau_{ia} = A$ and $\tau_{ib} = B$. Here we use the

notation that location $j = a$ has chemical A and location $j = b$ has chemical B. For each site having this chemical pair, write the pairs of dingo visits to the site as $Y_{it} = (Y_{iat}, Y_{ibt})$.

A simple bootstrap sample would simply ignore all spatial and temporal features of the data by sampling with replacement from the original dataset.

A slightly more complex bootstrap sample would still ignore the spatial and temporal dependencies in the data, but incorporate information about the design. For instance, this could mean sampling with replacement from the original data partitioned into site-location tuples according to the chemical applied there. Further complexity could be introduced by partitioning original data into site-location tuples according to the *pair of chemicals* applied there.

Estimates of chemical attractiveness are based on estimates of the probability of dingo presence which already incorporate spatial information: by blocking over groups of adjacent sites, or by smoothing over smaller groups of adjacent sites using binomial weights. So there is little need to alter the bootstrap sample to accommodate the spatial dependencies in the data.

The 'repeated measures' feature of the data, however, is ignored in the analysis, which assumes that observations at different times are effective replicates. We can capture this time dependency in our bootstrap samples by incorporating some time information in our sampling methodology. We choose a simple Markov Chain model of order one.

B.1.2 Markov chain bootstrap sampling method

We assume that the $\{Y_{it}\}$ follow a simple Markov Chain model where the one-step transition probability is constant over all sites with this chemical pair, and is given by:

$$P_{y_1}^{(AB)}(y_2) = P\left\{Y_{i,t} = y_2 | Y_{i,t-1} = y_1; \quad \tau_i = (A, B)\right\}.$$

Here $y_{ijt} \in \{0, 1\}$ so $y, y_1, y_2 \in \{00, 01, 10, 11\}$.

Thus we assume that over T days, the observed dingo visits to chemical pair (AB) located at site i have joint likelihood

$$P^{(AB)}(y_i) = P^{(AB)}(y_{i,1}) \prod_{t=2}^{T} P_{Y_{i,t-1}}^{(AB)}(y_{i,t}).$$

For the first time-point $t = 1$, we assume that the marginal probability is also constant over all sites having this chemical pair and is given by:

$$P^{(AB)}(y) = P\left\{Y_{i,t} = y; \quad \tau_i = (A, B)\right\}.$$

We can estimate these probabilities from the empirical distributions of our bootstrap samples. The marginal probability can be estimated from the empirical frequencies of dingo visits, computed over all sites having a certain chemical pair.

$$f^{(AB)}(y) = \sum_{i:\tau_i=(A,B)} \sum_{t=1}^{T} \mathcal{I} Y_{it} = y \tag{B.1}$$

Recall that on a given day, dingo visits were observed at 135 sites. There are $\binom{6}{2} = 15$ different chemical pairs. Since the design was balanced over chemical pairs, of these 135

sites, 9 sites have a given chemical pair (AB), We have collected observations over $T = 7$ days. So the estimated marginal probability is

$$\widehat{P}^{(AB)}(y) = \frac{f^{(AB)}(y)}{7 \times 9}. \tag{B.2}$$

The transition probabilities can similarly be estimated from the ratio of the empirical joint frequencies and the marginal frequencies of dingo visits, computed over all site-locations having a certain chemical pair. The joint frequencies are given by:

$$f^{(AB)}(y_2, y_1) = \sum_{i:\tau_i=(A,B)} \sum_{t=2}^{T} \mathcal{I} Y_{i,t} = y_2 \quad \text{and} \quad Y_{i,t-1} = y_1.$$

So the estimated transition probabilities are in fact conditional probabilities:

$$\widehat{P}^{(AB)}_{y_1}(y_2) = \frac{f^{(AB)}(y_2, y_1)}{f^{(AB)}(y_1)}. \tag{B.3}$$

Thus the bootstrap samples for sites with chemical pair $\tau_i = (A, B)$ may be generated according to:

$$Y^*_{it} = (Y^*_{iat}, Y^*_{ibt}) \sim \begin{cases} \widehat{P}^{(AB)}, & t = 1 \\ \widehat{P}^{(AB)}_{Y_{i,t-1}}, & t = 2, 3, \ldots, T \end{cases} \tag{B.4}$$

An iterative simulation progressing over days will produce a bootstrap sample.

Very occasionally, using this bootstrap sampling mechanism (Equation B.4), it is possible to obtain absolutely no visits to a pair of chemicals at all possible sites for a particular bootstrap sample, especially for chemical pairs with low visit probability. These samples were redrawn.

B.1.3 Results

For the dingo experiment, the bootstrap sampling distributions $\widehat{P}^{(AB)}$ and $\widehat{P}^{(AB)}_{Y_{i,t-1}}$ were calculated from the original data as specified in Equations (B.2) and (B.3). In Section B.1.3 we give the empirical distribution functions used to generate the bootstrap samples according to the regime specified in equation B.4.

Before collecting estimates from bootstrap samples, it was necessary to devise a general stopping criterion for the EM algorithm used to compute these estimates, which would ensure convergence (of some minimal level) for all bootstrap samples. The stopping criterion depended on the type of estimates used for the probability of dingo presence. Block estimates allowed the EM algorithm to converge much faster (within 200 iterations) than the smoothed estimates (requiring 10,000 iterations to ensure convergence to within 3 decimal places.) Details are given in Section B.1.3.

Once bootstrap samples of the estimates were obtained, their distribution can be investigated (Section B.1.3), and descriptive statistics can be computed (Section B.1.3).

Empirical distribution functions

Table B.1 shows the estimated marginal probabilities (Column 6) of dingo presence (Column 4) at pairs of chemicals (Columns 2 and 3). (Column 5 is not discussed here: see

Section B.2). For example, for chemicals 1 and 2, the estimated probability of no visit (00) to either chemical is 0.857.

For the given chemical pair, the estimated conditional probabilities for dingo presence at time $t+1$ given time t (Columns 7 to 10) add up to the estimated marginal probability of dingo presence at time t (Column 6), *c.f.* Equation (B.3). For example, for chemicals 1 and 2, the estimated probability of no visit (00) to either chemical on a certain day, assuming that only chemical 1 was visited the previous day, is 0.875. The probability that only chemical 1 was visited on any day, however, is only 0.127.

As can be seen from Table B.1, the highest estimated probabilities for any pair of chemicals is that no dingoes visited either chemical at the site, on any day.

Table B.1: Empirical distribution functions used to construct bootstrap samples. The first three columns give the chemical combination A and B at locations a and b from sites $\{i : \tau_i = (AB)\}$ which are involved in computing the marginal and conditional distributions of dingo visits. The fourth and fifth columns give the first day's data-based $P_1^{(AB)}(y_1)$ and complete-data based $P^{(AB)}(y_1)$ marginal probabilties of a dingo visit $y_1 \in \{00, 01, 10, 11\}$ at sites having chemical pair AB. The last four columns give the Markov chain transition probabilities of observing dingo visits to chemical pair AB on day $t+1$ given dingo visits to same chemical pair on day t, for $t = 1, \ldots, T$.

	Chemical		**Time t**			**Time $t+1$: $\mathbf{P^{(AB)}(y_2|y_1)}$**			
	A	B	y_1	$P_1^{(AB)}(y_1)$	$P^{(AB)}(y_1)$	$y_2 = 00$	$y_2 = 01$	$y_2 = 10$	$y_2 = 11$
1	1	2	00	0.778	0.857	0.867	0.000	0.111	0.022
2	1	2	01	0.000	0.000	0.000	0.000	0.000	0.000
3	1	2	10	0.222	0.127	0.875	0.000	0.125	0.000
4	1	2	11	0.000	0.016	1.000	0.000	0.000	0.000
5	1	3	00	1.000	0.937	0.941	0.000	0.039	0.020
6	1	3	01	0.000	0.016	1.000	0.000	0.000	0.000
7	1	3	10	0.000	0.032	0.000	1.000	0.000	0.000
8	1	3	11	0.000	0.016	1.000	0.000	0.000	0.000
9	1	4	00	0.667	0.810	0.837	0.047	0.047	0.070
10	1	4	01	0.111	0.048	1.000	0.000	0.000	0.000
11	1	4	10	0.000	0.063	1.000	0.000	0.000	0.000
12	1	4	11	0.222	0.079	0.600	0.000	0.400	0.000
13	1	5	00	0.667	0.810	0.886	0.023	0.045	0.045
14	1	5	01	0.111	0.048	0.667	0.000	0.333	0.000
15	1	5	10	0.000	0.063	0.500	0.000	0.500	0.000
16	1	5	11	0.222	0.079	0.600	0.200	0.000	0.200
17	1	6	00	0.778	0.762	0.780	0.122	0.073	0.024
18	1	6	01	0.000	0.111	0.667	0.167	0.167	0.000
19	1	6	10	0.111	0.095	0.800	0.200	0.000	0.000
20	1	6	11	0.111	0.032	0.500	0.000	0.500	0.000
21	2	3	00	1.000	0.952	0.941	0.039	0.020	0.000

continued on next page

Table B.1: (continued from previous page)

	Chemical		Time t			Time $t+1$: $\mathbf{P^{(AB)}(y_2|y_1)}$			
	A	B	y_1	$P_1^{(AB)}(y_1)$	$P^{(AB)}(y_1)$	$y_2 = 00$	$y_2 = 01$	$y_2 = 10$	$y_2 = 11$
22	2	3	01	0.000	0.032	1.000	0.000	0.000	0.000
23	2	3	10	0.000	0.016	1.000	0.000	0.000	0.000
24	2	3	11	0.000	0.000	0.000	0.000	0.000	0.000
25	2	4	00	0.889	0.889	0.875	0.125	0.000	0.000
26	2	4	01	0.000	0.095	1.000	0.000	0.000	0.000
27	2	4	10	0.000	0.000	0.000	0.000	0.000	0.000
28	2	4	11	0.111	0.016	1.000	0.000	0.000	0.000
29	2	5	00	0.778	0.873	0.957	0.043	0.000	0.000
30	2	5	01	0.222	0.095	0.600	0.200	0.200	0.000
31	2	5	10	0.000	0.032	0.000	0.500	0.500	0.000
32	2	5	11	0.000	0.000	0.000	0.000	0.000	0.000
33	2	6	00	1.000	1.000	1.000	0.000	0.000	0.000
34	2	6	01	0.000	0.000	0.000	0.000	0.000	0.000
35	2	6	10	0.000	0.000	0.000	0.000	0.000	0.000
36	2	6	11	0.000	0.000	0.000	0.000	0.000	0.000
37	3	4	00	1.000	0.857	0.870	0.109	0.000	0.022
38	3	4	01	0.000	0.095	0.500	0.167	0.167	0.167
39	3	4	10	0.000	0.016	0.000	0.000	0.000	0.000
40	3	4	11	0.000	0.032	1.000	0.000	0.000	0.000
41	3	5	00	0.889	0.921	0.918	0.061	0.020	0.000
42	3	5	01	0.111	0.063	1.000	0.000	0.000	0.000
43	3	5	10	0.000	0.016	1.000	0.000	0.000	0.000
44	3	5	11	0.000	0.000	0.000	0.000	0.000	0.000
45	3	6	00	0.889	0.952	0.962	0.038	0.000	0.000
46	3	6	01	0.000	0.032	1.000	0.000	0.000	0.000
47	3	6	10	0.111	0.016	1.000	0.000	0.000	0.000
48	3	6	11	0.000	0.000	0.000	0.000	0.000	0.000
49	4	5	00	0.889	0.921	0.959	0.020	0.020	0.000
50	4	5	01	0.000	0.032	0.500	0.000	0.500	0.000
51	4	5	10	0.111	0.048	0.667	0.333	0.000	0.000
52	4	5	11	0.000	0.000	0.000	0.000	0.000	0.000
53	4	6	00	0.778	0.905	0.940	0.000	0.040	0.020
54	4	6	01	0.000	0.016	1.000	0.000	0.000	0.000
55	4	6	10	0.111	0.048	1.000	0.000	0.000	0.000
56	4	6	11	0.111	0.032	0.000	1.000	0.000	0.000
57	5	6	00	0.889	0.873	0.870	0.022	0.065	0.043
58	5	6	01	0.000	0.032	1.000	0.000	0.000	0.000
59	5	6	10	0.000	0.048	0.667	0.333	0.000	0.000
60	5	6	11	0.111	0.048	1.000	0.000	0.000	0.000

EM stopping criterion

Diagnostic plots were used to determine an EM stopping criterion. The simplest plot shows estimates of chemical attractiveness *vs* EM iterations, and can be used to approximately

locate time of convergence by eye, as well as check the pattern of convergence. It is well-known that EM has been observed to exhibit cyclic tendencies.

In some cases the convergence appears to be slowing down, but still increasing, so locating the EM time of convergence by eye is fairly arbitrary and a more specific stopping criterion is required, *e.g.* percentage change in estimates over EM iterations.

Block estimates of probability of dingo presence We decided that only 200 iterations would be sufficient to ensure convergence of estimates, over a wide range of block sizes. Figure 1 for blocksize 15 is typical of different bootstrap samples for all sensible blocksizes, and demonstrates that convergence has occurred well before 200 iterations.

(Binomially) smoothed estimates of probability of dingo presence Figure 2 shows the behaviour of estimates of chemical attractiveness $\{\widehat{q}_k\}$, computed by smoothing the estimates of probability of dingo presence over 2 adjacent sites in both directions, for every 10th iteration (due to storage restrictions) in an EM algorithm which was allowed to cycle for 100,000 iterations. (The plot only focusses on the first 40,000 iterations.) It appears from this plot that the estimates are converging after 5,000 iterations. Strange behaviour of the estimates is evident as convergence is approached: the estimates apear to cycle back in quantum jumps to previous values, cycling less and less frequently (with similar jumps) as convergence is finally reached. On closer inspection of the percentage change in the estimates, (Table B.2) we see that the algorithm achieves convergence to at most 0.01%, or approximately 4 decimal places, in at least 30,000 iterations in some cases! Similar behaviour was demonstrated for smoothing over 1 or 3 adjacent sites in both directions.

Table B.2: Number of EM iterations required to ensure that the given % change was achieved for each chemical effectiveness parameter $\{q_k; k = 1, \ldots, 6\}$.

% change	Chemical effectiveness parameter					
	q_1	q_2	q_3	q_4	q_5	q_6
0.15	410	1710	2080	1230	870	940
0.14	420	1710	2490	1250	1090	1190
0.13	540	1710	3130	1400	1090	1290
0.12	540	1710	3980	1580	1420	1290
0.11	740	1850	5100	1580	1420	1640
0.10	740	2880	5100	1790	1760	1640
0.09	1110	4700	5100	1990	1760	2160
0.08	1110	10270	5100	1990	2180	2160
0.07	1570	24970	5100	2240	2180	2750
0.06	1570	24970	5100	2630	2890	2860
0.05	2450	24970	7620	2630	2890	3820
0.04	2450	24970	29300	3060	3740	3890
0.03	4130	24970	29300	3750	3810	5810
0.02	4130	24970	29300	5090	5650	5810
0.01	27070	24970	29300	28940	32520	34240

The unusual cyclic pattern was again seen across the percentage change over EM iterations. Over EM time, the percentage change gradually decreases, but in seemingly quantum

leaps. Using the information on percentage change, we decided that 10,000 iterations of the algorithm would be sufficient to ensure convergence of estimates, to 0.1%, or approximately 3 decimal places, provided a check was made to ensure that the final EM iteration was not one of the retrograde quantum cycles. This is *vastly* greater than the number of iterations required for convergence when block estimates of the probability of dingo presence were used, and suggests that the numerical process is not well behaved.

Distribution of bootstrap estimates

Block estimates of probability of dingo presence A typical distribution of bootstrap estimates obtained for block estimates of the probability of dingo presence indicates that, over bootstrap samples having the same simple Markov Chain properties, the larger estimates are approximately normally distributed, whilst the smaller estimates are much 'peakier' than the closest normal. (In this particular case, blocksize is 15. Density estimates use 20 points and Gaussian windows.) Slight amounts of skewness are apparent for $\widehat{q}_2$.

A typical distribution of the *logged ratios* of bootstrap estimates (for all except the first), again using block estimates, showed that although the skewness has been improved, the 'peakiness' appears to have increased for the last three estimates ($\widehat{q}_4, \widehat{q}_5, \widehat{q}_6$.) (The particular case is again for blocksize 15.) No real benefit is gained by the transformation in terms of improving the distribution of estimates.

(Binomially) smoothed estimates of probability of dingo presence The distribution of estimates when probability of dingo presence is smoothed over 2 adjacent sites to both sides are not very similar in shape to those distributions for blocked estimates above. The location of all estimates is much higher; and the distribution for estimate $\widehat{q}_1$ appears to be slightly 'peakier', whilst estimates $\widehat{q}_2, \widehat{q}_3$ have less 'peaky' distributions.

After taking the log ratios of estimates, the distributions all appear to be much peakier, with the possible exception of q_2.

Similar behaviour was noted for smoothing over 1 or 3 adjacent sites in both directions.

Descriptive statistics of bootstrap estimates

Mean, standard deviation and standard error In Table B.3 below, the results observed in the plots described above are quantified more precisely. Mean estimates of $\widehat{q}_k$ are approximately 0.45, 0.06, 0.15, 0.42, 0.30, and 0.28; and vary slightly (up to .03) for blocked estimates of the probability of dingo presence. When smoothing is used, the estimates $\widehat{q}_k$ are 0.22, 0.08, 0.22, 0.32, 0.30, and 0.32 higher respectively, and also vary slightly (up to .02).

As observed from the plots, the standard deviation is similar for blocked and smoothed cases, for estimates $\widehat{q}_k, k = 4, 5, 6$. There is slightly less variation in smoothed compared to block cases of the first estimate $\widehat{q}_1$; whilst there is almost double the standard deviation for smoothed compared to block cases of the second and third estimates.

The standard errors of the mean show a similar pattern to the standard deviation, although the differences are smaller (due to the scale) and approximate around 0.002–0.003.

Quantiles (Interval estimates) The empirical bootstrapped 90% confidence intervals for the estimates are simply obtained via the 5th and 95th percentiles of the bootstrapped estimates.

Table B.3: Descriptive statistics of bootstrapped estimates of chemical effectiveness for smoothed and blocked estimates of the probability of dingo presence, and for various block sizes.

smooth/ block	**block size**	**chemical attractiveness estimate** $\widehat{q}_1$	$\widehat{q}_2$	$\widehat{q}_3$	$\widehat{q}_4$	$\widehat{q}_5$	$\widehat{q}_6$
Mean							
B	5	0.470497	0.064299	0.147625	0.442314	0.316253	0.285791
B	9	0.452675	0.061998	0.145822	0.415871	0.302743	0.275615
B	15	0.446659	0.061931	0.145689	0.415218	0.301845	0.274738
B	27	0.444000	0.062205	0.146430	0.416161	0.302226	0.276223
S	1	0.674811	0.148460	0.335902	0.740735	0.576165	0.568131
S	2	0.695260	0.168339	0.372675	0.766639	0.608904	0.608001
S	3	0.699926	0.173656	0.381986	0.772276	0.616447	0.617253
Standard deviation							
B	5	0.078794	0.030748	0.049529	0.092089	0.075224	0.069260
B	9	0.082332	0.030420	0.051289	0.096431	0.078449	0.071490
B	15	0.084990	0.030709	0.052164	0.099217	0.080568	0.073144
B	27	0.085547	0.030951	0.052674	0.099980	0.081178	0.074238
S	1	0.067454	0.065116	0.090754	0.073566	0.087629	0.090737
S	2	0.065524	0.072704	0.096468	0.068977	0.086777	0.089187
S	3	0.065014	0.074662	0.097648	0.067870	0.086410	0.088543
Standard error of the Mean							
B	5	0.002492	0.000972	0.001566	0.002912	0.002379	0.002190
B	9	0.002604	0.000962	0.001622	0.003049	0.002481	0.002261
B	15	0.002688	0.000971	0.001650	0.003138	0.002548	0.002313
B	27	0.002705	0.000979	0.001666	0.003162	0.002567	0.002348
S	1	0.002133	0.002059	0.002870	0.002326	0.002771	0.002869
S	2	0.002072	0.002299	0.003051	0.002181	0.002744	0.002820
S	3	0.002056	0.002361	0.003088	0.002146	0.002733	0.002800

Table B.4: 90% Interval estimates of bootstrapped estimates of chemical effectiveness for smoothed and blocked estimates of the probability of dingo presence, and for various block sizes.

smooth/ block	block size	chemical attractiveness estimate					
		$\widehat{q}_1$	$\widehat{q}_2$	$\widehat{q}_3$	$\widehat{q}_4$	$\widehat{q}_5$	$\widehat{q}_6$
Lower 90% interval estimate							
B	5	0.3491	0.0179	0.0718	0.3087	0.2047	0.1845
B	9	0.3225	0.0178	0.0703	0.2711	0.1886	0.1707
B	15	0.3094	0.0181	0.0682	0.2626	0.1809	0.1671
B	27	0.3062	0.0180	0.0676	0.2652	0.1821	0.1679
S	1	0.5622	0.0436	0.1935	0.6208	0.4408	0.4138
S	2	0.5870	0.0521	0.2182	0.6555	0.4743	0.4557
S	3	0.5922	0.0546	0.2244	0.6635	0.4820	0.4652
Upper 90% interval estimate							
B	5	0.6014	0.1210	0.2359	0.6029	0.4502	0.4087
B	9	0.5876	0.1164	0.2413	0.5819	0.4445	0.4036
B	15	0.5866	0.1166	0.2442	0.5860	0.4412	0.4064
B	27	0.5849	0.1178	0.2455	0.5889	0.4469	0.4090
S	1	0.7805	0.2647	0.4864	0.8601	0.7178	0.7132
S	2	0.7990	0.2976	0.5364	0.8775	0.7478	0.7483
S	3	0.8029	0.3076	0.5458	0.8812	0.7543	0.7568

In Table B.4, we can see a similar pattern for the lower and upper interval estimates, as compared to the pattern of the mean and standard deviation. The estimates obtained by using blocked or smoothed estimates of the probability of dingo presence are very different, reflecting the difference in the mean estimates for these two cases. The length of the interval appears slightly larger for the block case, which also reflects the larger standard deviations in this case.

The difference between chemical attractiveness estimate q_1 and q_4 is more marked in the smoothed case: respective 90% intervals are $(.58, .80)$ and $(.66, .88)$ for the smoothed case and $(.32, .59)$ and $(.27, .59)$ for the blocked case. In either case, the interval estimates of q_5 and q_6 are barely separated. In the blocked case, the interval estimate is approximately (0.2,0.40–0.45) and in the smoothed case, it is approximately (0.45,0.75).

The length of the interval estimate of q_2 is three time larger in the smoothed case, and the interval limits of q_3 are more than double the size in the smoothed case. In the blocked case, the interval estimate of q_2 is about $(0.02, 0.12)$, and in the smoothed case it is about $(0.05, 0.30)$. In the blocked case, the limits of q_3 are about $(0.07, 0.24)$, and in the smoothed case these are about $(0.21, 0.54)$.

B.2 Alternative bootstrap sampling strategies

The results above were obtained by drawing bootstrap samples of dingo visits to particular chemical pairs over all site-day combinations, using a Markov Chain approach (Strategy I: see Section B.1.2.) That is, samples for visits to chemical pairs on the first day were drawn from the *overall* empirical marginal distribution of visits to those chemical pairs observed over all day-site tuples. Then samples for visits on subsequent days were drawn from the empirical conditional distribution of visits to chemical pairs given the previous day's visits to the same chemical pairs, again observed over all day-site tuples for that chemical pair. This empirical conditional distribution was computed by considering all pairs of previous and next days—6 altogether—for all chemical pairs.

One alternative (Strategy II) is to obtain bootstrap samples for the first day from the *first day's* empirical marginal distribution of visits, instead of from the *overall* empirical marginal distribution, obtained by aggregating observed frequencies over all days. It can be argued that obtaining the initial distribution for visits from just the first day's data makes more sense.

Another alternative (Strategy III) is to obtain samples for visits on the first day from the *first day's actual data* instead of from the *first day's* empirical marginal distribution, which combines all the first day's information. It can be argued that a good initial starting point for day 1 of the Markov Chain can be given by the sample that we observed.

B.2.1 Strategy II: First day's bootstrap sample generated from empirical distribution of first day's data

For the first scenario, the marginal probability can be estimated from the empirical frequencies of dingo visits, *only on the first day*, computed over all site-locations having a certain chemical pair. (Compare this to the marginal estimated from all days visits in equation (B.1).)

$$f_1^{(AB)}(y) = \sum_{i:\tau_i=(A,B)} \mathcal{I}Y_{it} = y \tag{B.5}$$

And so the bootstrap sampling distribution for the first day's visits becomes:

$$\widehat{P_1}^{(AB)}(y) = \frac{f_1^{(AB)}(y)}{9}. \tag{B.6}$$

Thus the bootstrap samples for sites with chemical pair $\tau_i = (A, B)$ may be generated according to:

$$Y_{it}^* = (Y_{iat}^*, Y_{ibt}^*) \sim \begin{cases} \widehat{P_1}^{(AB)}, & t = 1 \\ \widehat{P}_{Y_{i,t-1}}^{(AB)}, & t = 2, 3, \ldots, T \end{cases} \tag{B.7}$$

This replaces equation (B.4) given previously, and only differs in the sampling distribution for the case $t = 1$.

B.2.2 Strategy III: First day's bootstrap sample obtained from first day's observed data

As an alternative to sampling the first day's visits using the empirical marginal based on every day's data (equation (B.4)) or based on just the first day's data (equation (B.7)), we

can use the observed data:

$$Y^*_{it} = (Y^*_{iat}, Y^*_{ibt}) \begin{cases} = (y_{iat}, y_{ibt}), & t = 1 \\ \sim \widehat{P}^{(AB)}_{Y_{i,t-1}}, & t = 2, 3, \ldots, T \end{cases} \tag{B.8}$$

As before, every other day's visits are sampled from the empirical Markov Chain conditional distribution based on the previous day's visits.

B.2.3 Results: Strategy II

We focus on using the more stable "blocked" estimates for the probability of dingo presence, using blocksizes of 5, 9, and 15 sites.

The empirical marginal distribution for visits to chemical pairs was re-computed from just the first day's data, according to equation (B.7) above. Obviously, this marginal distribution is much more sparse, comprising many more zeroes than when data over all days was used.

Table B.5 includes some descriptive statistics on the chemical attractiveness estimates obtained from using $\widehat{P_1}^{(AB)}$ in equations B.5,B.6, and B.7, instead of $\widehat{P}^{(AB)}$ (see Table B.3).

There is little difference, (before the third significant digit), between the estimates obtained with these three different blocksizes for the blocked estimates of the probability of dingo presence.

The chemical attractiveness estimates are all a little higher now, *e.g.* for chemical 1, the estimate is approximately 0.51 with $\widehat{P_1}^{(AB)}$, instead of 0.45 obtained with $\widehat{P}^{(AB)}$. This can be explained by noticing that on the first day, the distribution of visits to chemical pairs was very different to that observed over all days. (See Table B.1.) In particular the first day's distribution appears to be less skewed towards the very low values, which would help explain why all estimates are slightly higher.

Since the bootstrap's first day's sample were restricted to a smaller set of possible values, from just the first day instead of all days, there was less variation in the bootstrap samples. Hence, as to be expected, the standard error in the chemical attractiveness estimates was lower with this strategy compared to the first strategy.

Similar patterns of distribution in the estimates are observed as before.

Table B.5: Descriptive statistics of bootstrapped estimates of chemical effectiveness for smoothed and blocked estimates of the probability of dingo presence, for various block sizes. Marginal distribution sampled for first day's bootstrap data was obtained from first day's data only ($\widehat{P_1}^{(AB)}$.)

smooth/ block	**block size**	**chemical attractiveness estimate**					
		$\widehat{q}_1$	$\widehat{q}_2$	$\widehat{q}_3$	$\widehat{q}_4$	$\widehat{q}_5$	$\widehat{q}_6$
Mean							
B	5	0.521007	0.074887	0.146516	0.486952	0.389427	0.303288
B	9	0.510421	0.073957	0.148073	0.471948	0.381123	0.300685
B	15	0.507298	0.074395	0.149580	0.474524	0.382502	0.301028
Standard deviation							
B	5	0.079412	0.033142	0.051049	0.093251	0.082455	0.072288
B	9	0.081972	0.033107	0.052811	0.096779	0.084913	0.074239
B	15	0.083184	0.033412	0.053704	0.097708	0.085939	0.074734
Standard error of the Mean							
B	5	0.002511	0.001048	0.001614	0.002949	0.002607	0.002286
B	9	0.002592	0.001047	0.001670	0.003060	0.002685	0.002348
B	15	0.002630	0.001057	0.001698	0.003090	0.002718	0.002363

Table B.6: Descriptive statistics of bootstrapped estimates of chemical effectiveness for smoothed and blocked estimates of the probability of dingo presence, for various block sizes. The first day's bootstrapped data, at each site, was obtained from the observed first day's data, at each site. ($y^*_{i1} = \widehat{y_{i1}}$.)

smooth/ block	block size	chemical attractiveness estimate					
		$\widehat{q}_1$	$\widehat{q}_2$	$\widehat{q}_3$	$\widehat{q}_4$	$\widehat{q}_5$	$\widehat{q}_6$
Mean							
B	5	0.533169	0.077355	0.153571	0.491769	0.401576	0.313337
B	9	0.531626	0.074578	0.151888	0.483554	0.391247	0.312663
B	15	0.514214	0.074193	0.150572	0.470168	0.385387	0.306126
Standard deviation							
B	5	0.067726	0.030830	0.052419	0.082534	0.072455	0.066049
B	9	0.068156	0.029550	0.051311	0.080823	0.069762	0.063641
B	15	0.071413	0.030131	0.053257	0.088589	0.074174	0.066925
Standard error of the Mean							
B	5	0.002142	0.000975	0.001658	0.002610	0.002291	0.002089
B	9	0.002155	0.000934	0.001623	0.002556	0.002206	0.002013
B	15	0.002258	0.000953	0.001684	0.002801	0.002346	0.002116

B.2.4 Results: Strategy III

The bootstrap sample for the first day's visits to each site were not drawn from any random distribution, but instead were set to the original observed values at each site on the first day.

Since the bootstrap's first day's sample were fixed to correspond with the observed data, there was less variation in the bootstrap samples, than either of the less restricted bootstrap sampling schemes. Hence, as to be expected, the standard error in the chemical attractiveness estimates was lower with this strategy compared to both the other two strategies.

However, for each chemical, the bootstrapped mean estimates (over 1000 bootstrap simulations) were slightly higher than with Strategy II, and higher than with Strategy I. Note that substantially more sites were visited on Day 1 (27/128) and to a lesser degree Day 4 (24/128) compared to other days. Hence by focussing only on Day 1 data to generate bootstrap samples for Day 1, on average, we are increasing the overall number of visits in our bootstrap sample. In turn, this leads to higher absolute values of chemical attractiveness estimates, since more visits imply higher chemical attractiveness.

Furthermore, estimates would be higher using exact data instead of resampled data for Day 1 if the exact data promotes more visits on Day 2 via the Markov Chain model, compared to a random resampling of Day 1 data. This could be explained, for instance, if there is strong dependence between Day 1 and 2 visits, which cannot be accounted for fully by the chemical pair at the site, but instead to a large degree depends on the previous visit to the site.

Appendix C

MCMC diagnostics for *dingo* case study, Extension I

Table C.1: Extension I, Experiment 1: Distributional choices for MCMC construction for Dingo experiment

Parameter	Prior distribution	Proposal distribution	Starting values
Scenario 1			
$q = \exp \alpha$	Rotated uniform (0,1)[0,1] Truncated uniform $(0,1)[0,1]$	Truncated uniform $(q \pm 0.1)[0,1]$ Rotated uniform $(q \pm 0.1)[0,1]$	0.3
θ	Discrete uniform $[1 \ldots 48]$	Discrete NN (diag) Order 2	as for Prior
z_{st}		N/A	Bernoulli(0.2)
Scenario 2			
β	Truncated $N(0,1)[-10,+10]$	Truncated $N(\beta,1)[-10,+10]$	0.0
α	Truncated $N(0,3)[-10,+10]$	Truncated $N(\alpha,1)[-10,+10]$	0.1
θ	Discrete uniform $[1 \ldots 12]$	Discrete NN (diag) Order 2	as for Prior
z_{st}		N/A	Bernoulli(0.2)

Table C.2: Distributional choices for MCMC construction for experiment 2

Parameter	Prior distribution	Proposal distribution	Starting values
Scenario 1			
$q = \exp \alpha$	Rotated uniform (0,1)[0,1] Truncated uniform $(0,1)[0,1]$	Truncated uniform $(q \pm 0.1)[0,1]$ Rotated uniform $(q \pm 0.1)[0,1]$	0.3
θ	Discrete uniform $[1 \ldots 48]$	Discrete NN (diag) Order 2	as for Prior
X		N/A	0.2
Scenario 2			
β	Truncated $N(0,1)[-10,+10]$	Truncated $N(\beta,1)[-10,+10]$	0.0
α	Truncated $N(0,3)[-10,+10]$	Truncated $N(\beta,1)[-10,+10]$	0.1
θ	Discrete uniform $[1 \ldots 12]$	Discrete NN (diag) Order 2	as for Prior
X		N/A	0.2

Appendix D

MCMC diagnostics for *dingo* case study, Extension I, Experiment 2

I present here the MCMC diagnostics which support inference for the four-tier Extension I to the basic hierarchical model, as applied to the *dingo* case study. Estimation of the posterior distributions of parameters were the aim of inference.

Diagnostics are presented first for the spatio-temporal parameters θ which govern the presence/absence process z via the prior $p(z \mid \theta)$. Next diagnostics are presented for natural statistics from z given by $V_k(z)$, $k = 0, 1, 2$. This is more efficient than analyzing output from 839 individual $\tilde{z}_{st}$ chains, and the natural statistics are important in the model. Lastly the coefficients α for the explanatory variable x are explored via their counterpart q, due to the one-to-one relationship between α and q.

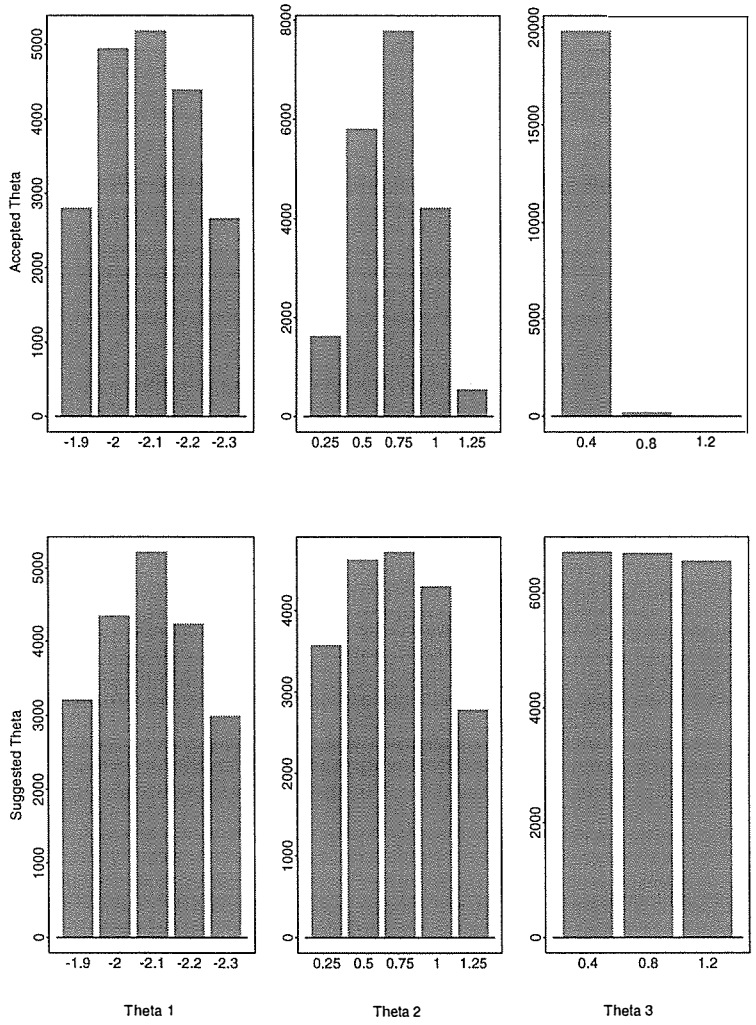


Figure D.1: Extension I, Experiment 2: Posterior distribution of θ values compared to distribution of proposed θ^* values

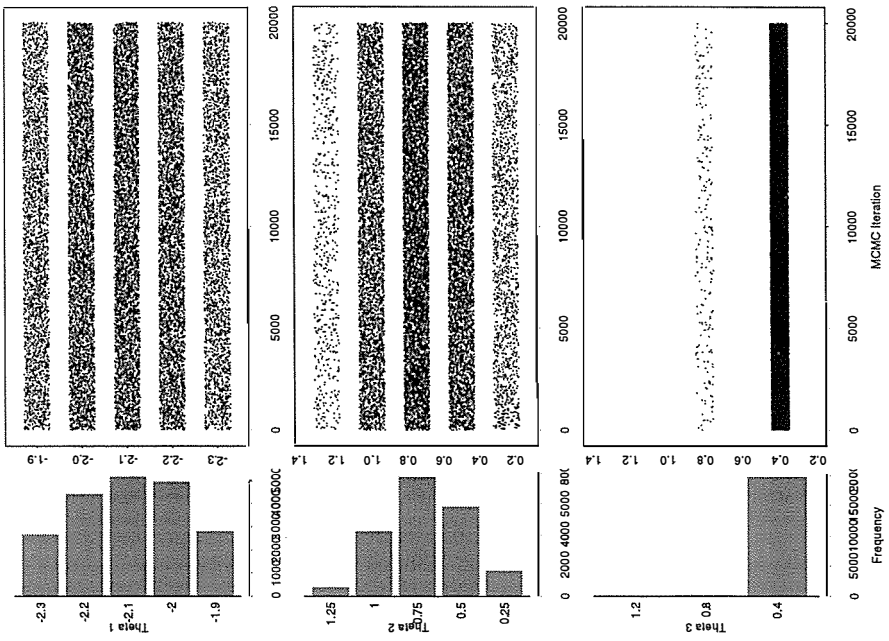


Figure D.2: Extension I, Experiment 2: Posterior distribution and MCMC time series of θ components. Note that the y-coordinate has been 'jittered' to improve the display.

Table D.1: Extension I, Experiment 2: MCMC descriptive statistics of posterior distribution simulations of the natural statistics for presence/absence $V(x)$.

Parameter	Sample Mean (Stdev)	SE (Naive TS Batch)	Lag 1 AC	Lag 1 Batch AC	50% Credible Interval
V_0	159 (13)	0.0946 0.0834 0.0919	0.0123	-0.0622	[149, 167]
V_1	44 (8.96)	0.0634 0.0574 0.0611	0.0099	-0.0254	[38,49]
V_2	38.1 (5.88)	0.0416 0.0364 0.0401	0.0057	-0.0670	[34,42]

Table D.2: Extension I, Experiment 2: MCMC Convergence Diagnostics for presence/absence natural statistics $V(x)$

Parameter	Geweke Z	Raftery-Lewis	Heidelberger-Welch Tests CVM* Stationarity		Halfwidth	
$V_0(x)$	0.664	1,4711	$\surd$	0.08	$\surd$	0.163
$V_1(x)$	0.983	2,5239	$\surd$	0.08	$\surd$	0.113
$V_2(x)$	1.030	2,6711	$\surd$	0.05	$\surd$	0.071

* CVM is an abbreviation for Cramer Von Mises statistic

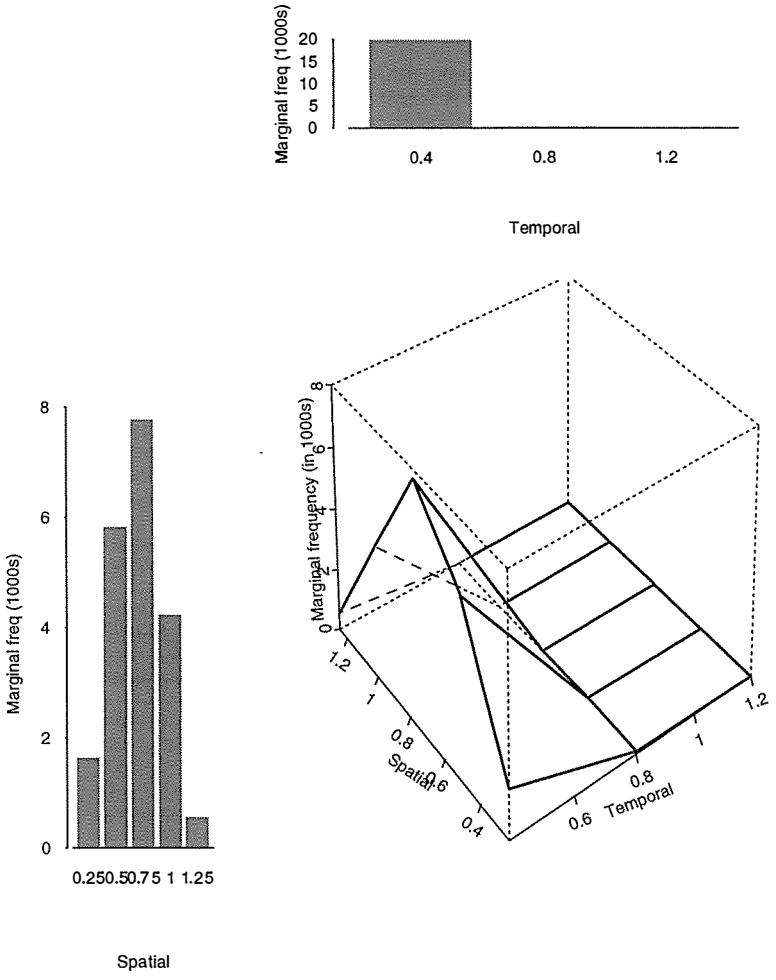


Figure D.3: Extension I, Experiment 2: Joint posterior distribution of (θ_1, θ_2), components of θ.

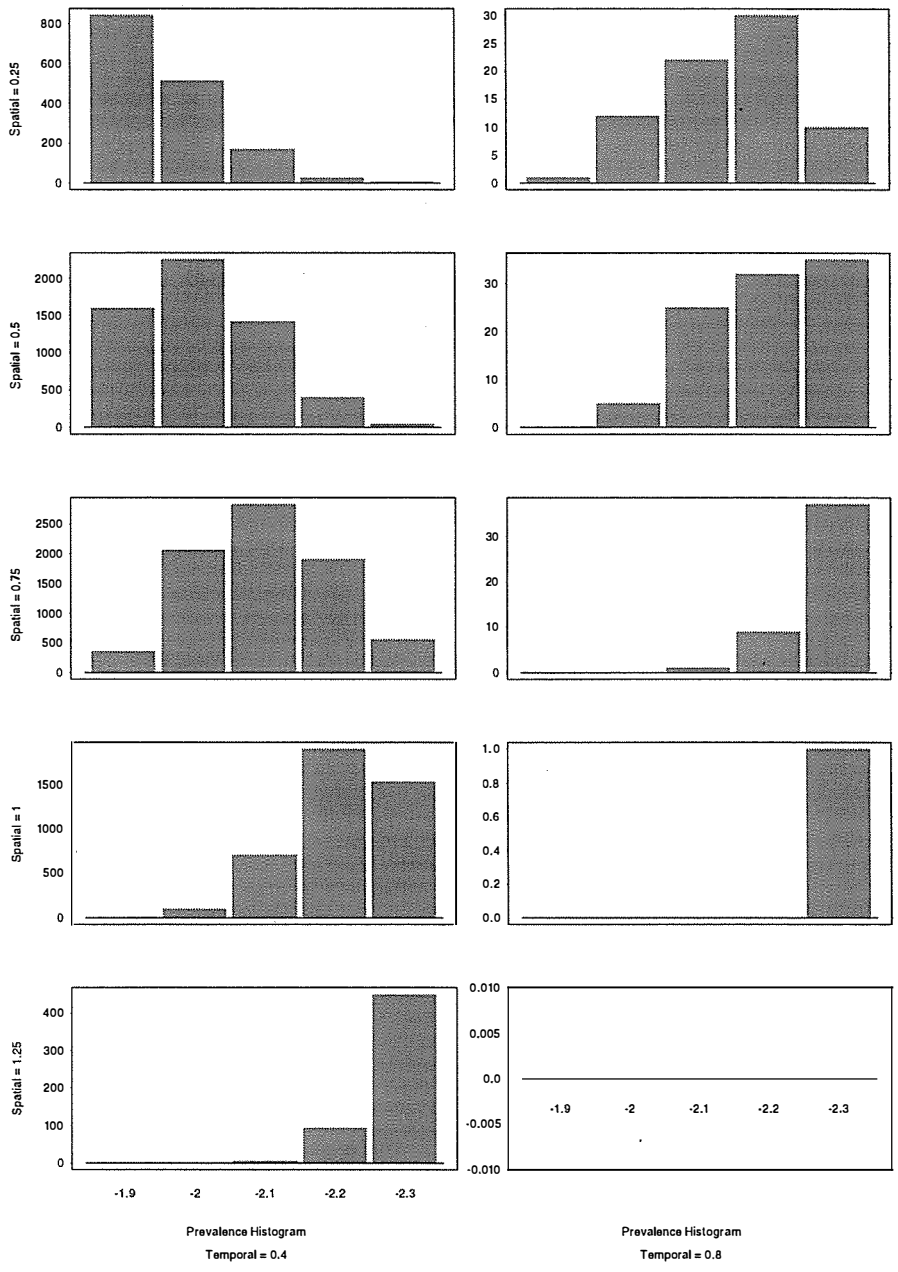


Figure D.4: Extension I, Experiment 2: Posterior distribution of θ_0 holding (θ_1, θ_2) constant.

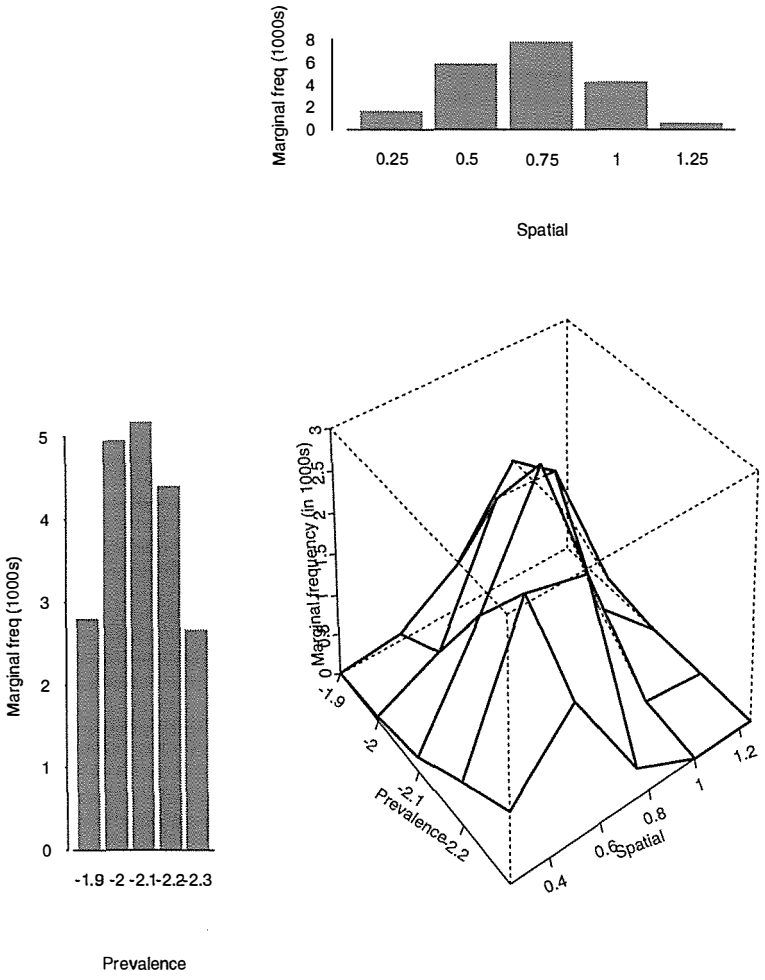


Figure D.5: Extension I, Experiment 2: Joint posterior distribution of (θ_0, θ_1), components of θ.

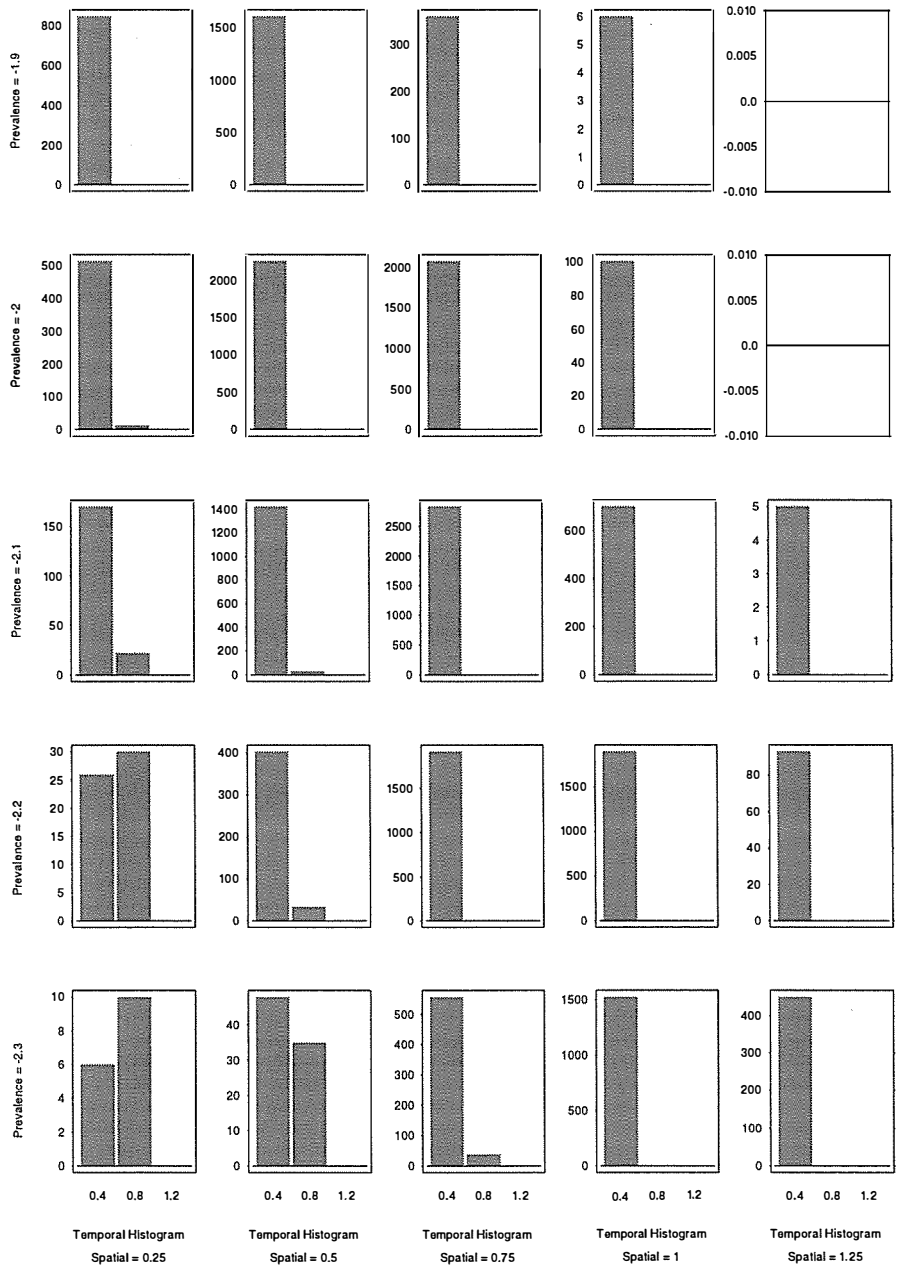


Figure D.6: Extension I, Experiment 2: Posterior distribution of θ_2 holding (θ_0, θ_1) constant.

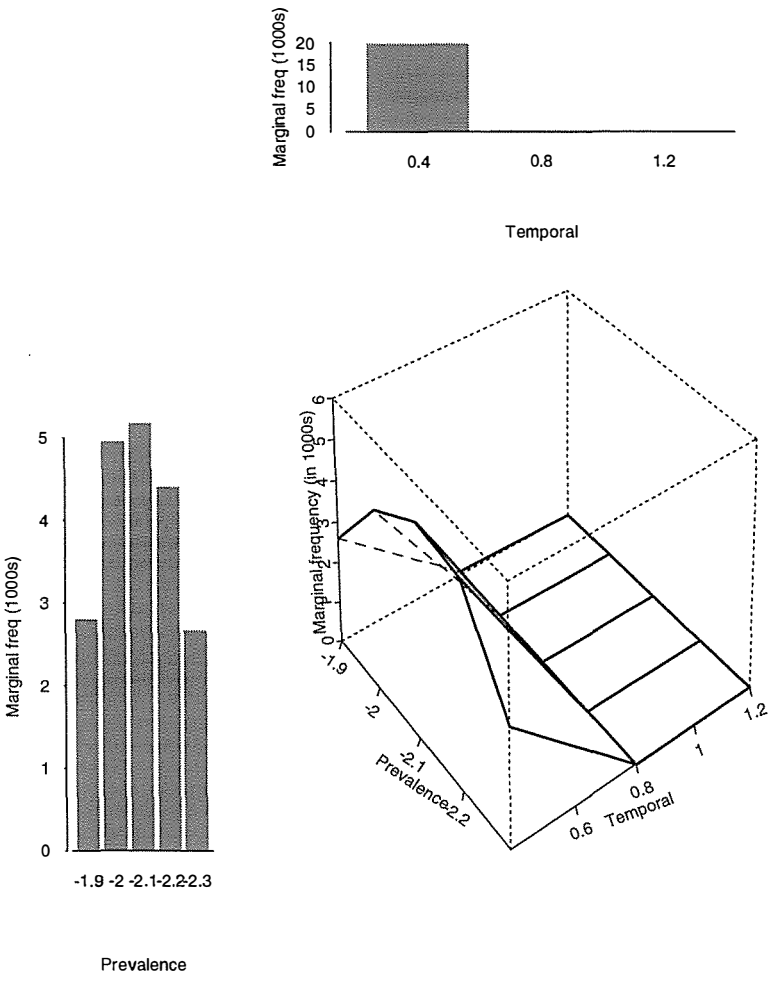


Figure D.7: Extension I, Experiment 2: Joint posterior distribution of (θ_0, θ_2), components of θ.

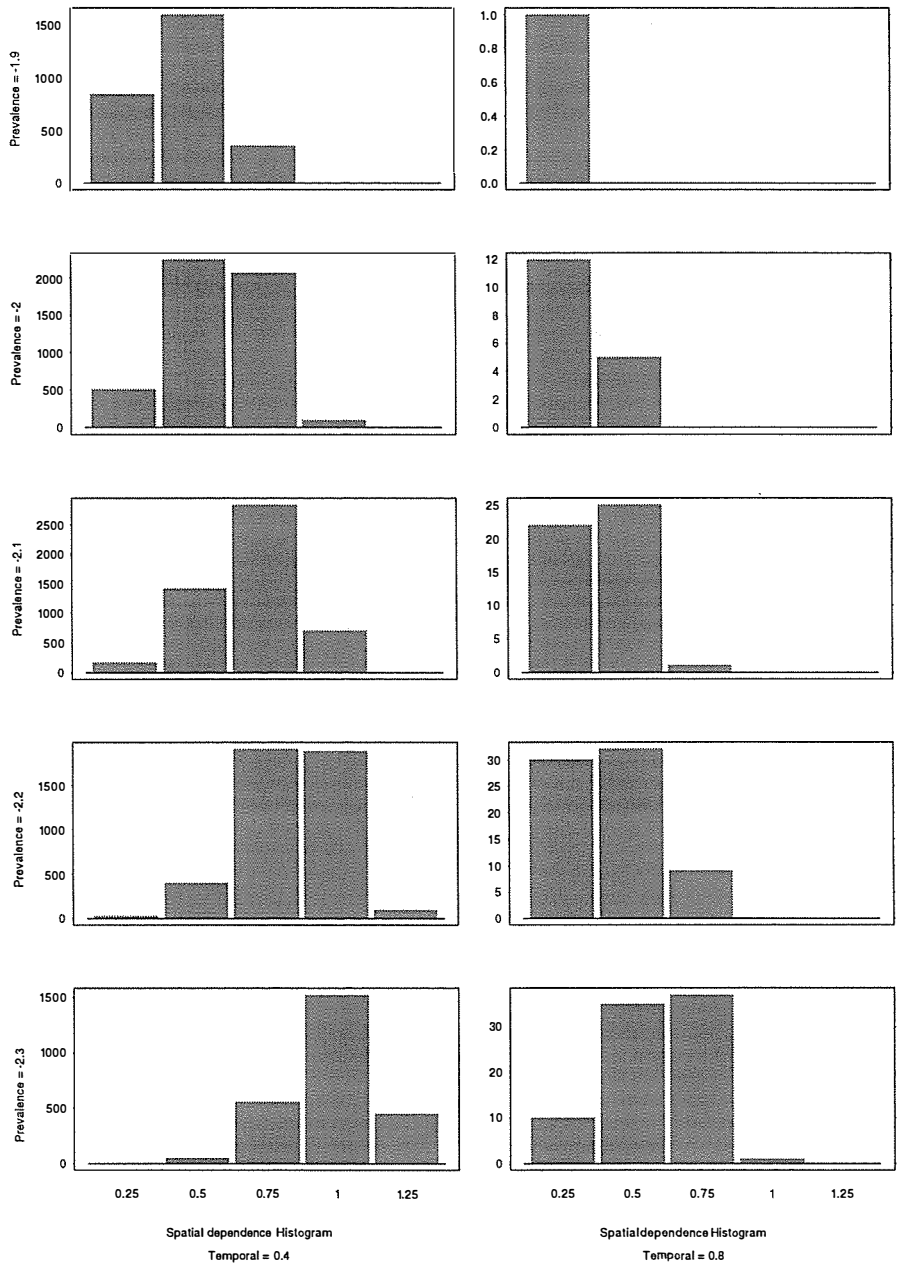


Figure D.8: Extension I, Experiment 2: Posterior distribution of θ_1 holding (θ_0, θ_2) constant

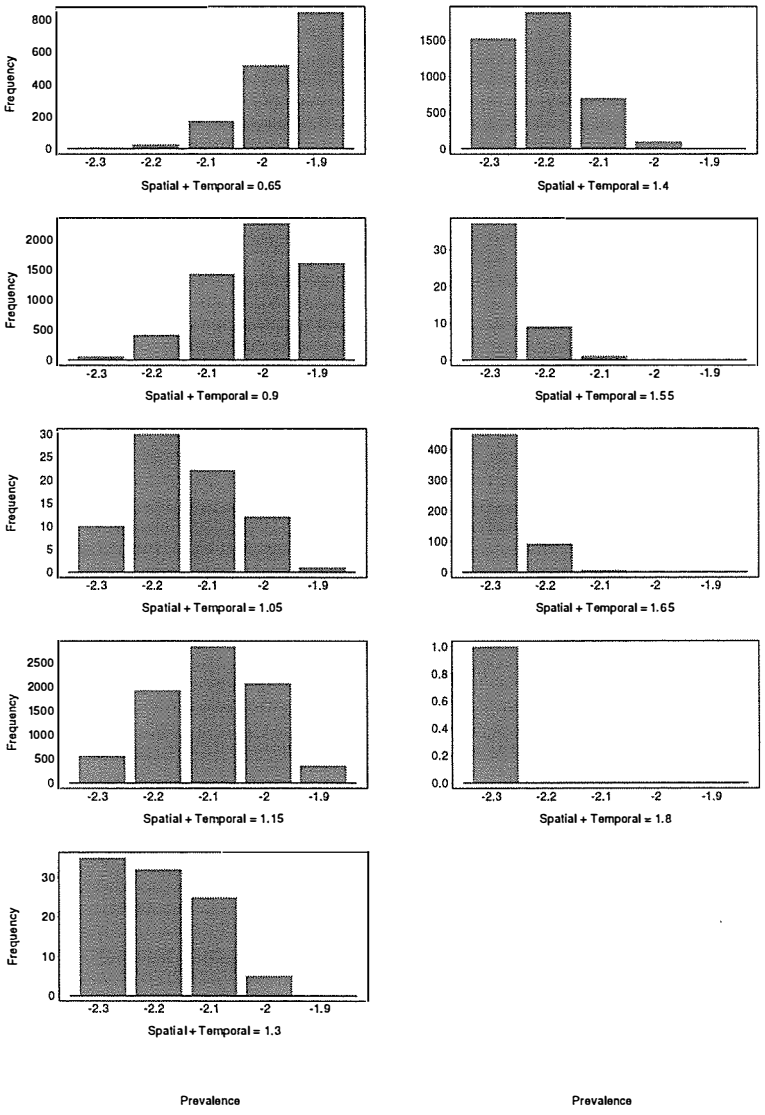


Figure D.9: Extension I, Experiment 2: Distribution of prevalence parameter θ_0, compared to overall dependence as measured by $\theta_1 + \theta_2$.

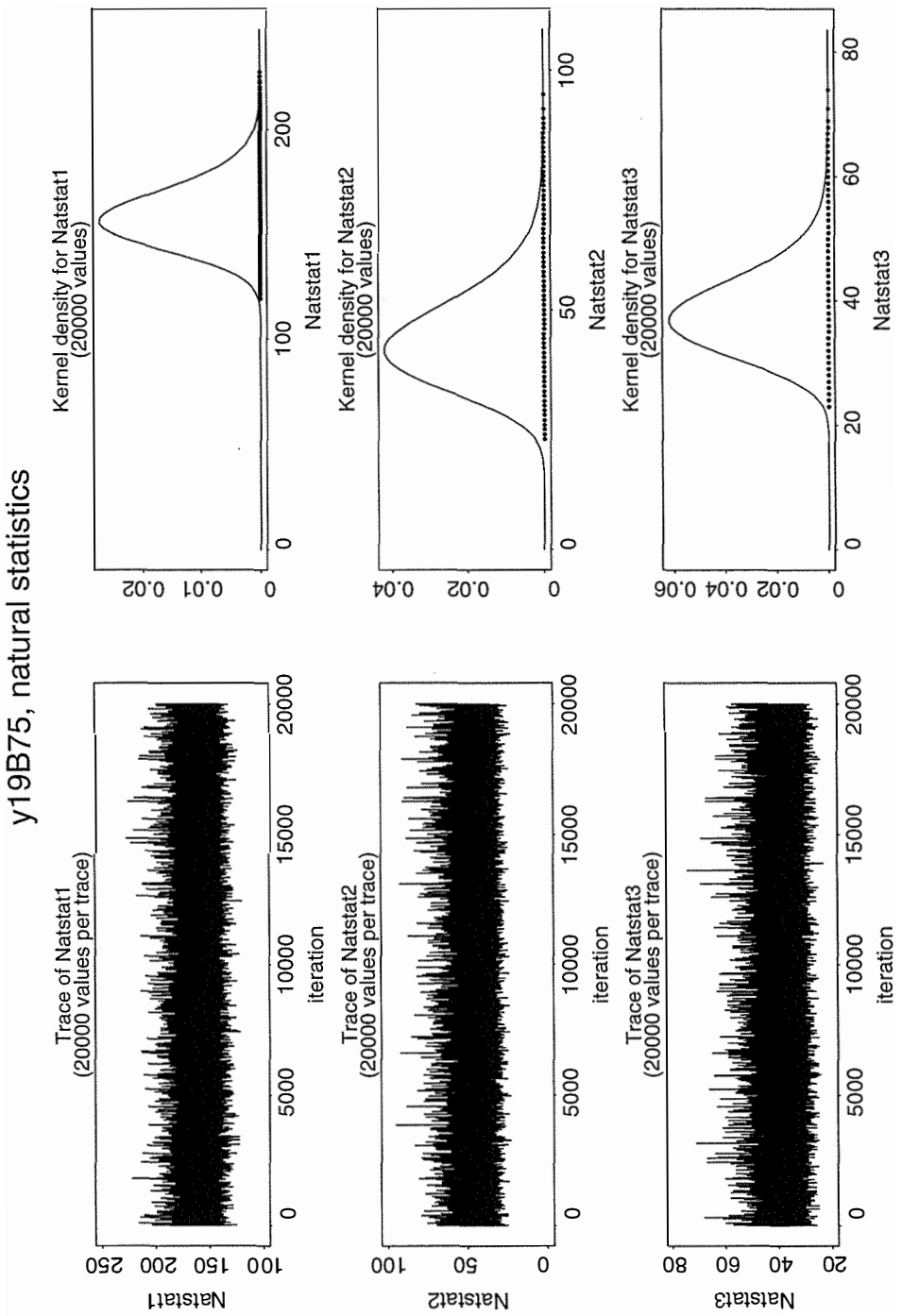


Figure D.10: Extension I, Experiment 2: MCMC Trace of the posterior distribution of natural statistics of presence/absence $V(x)$

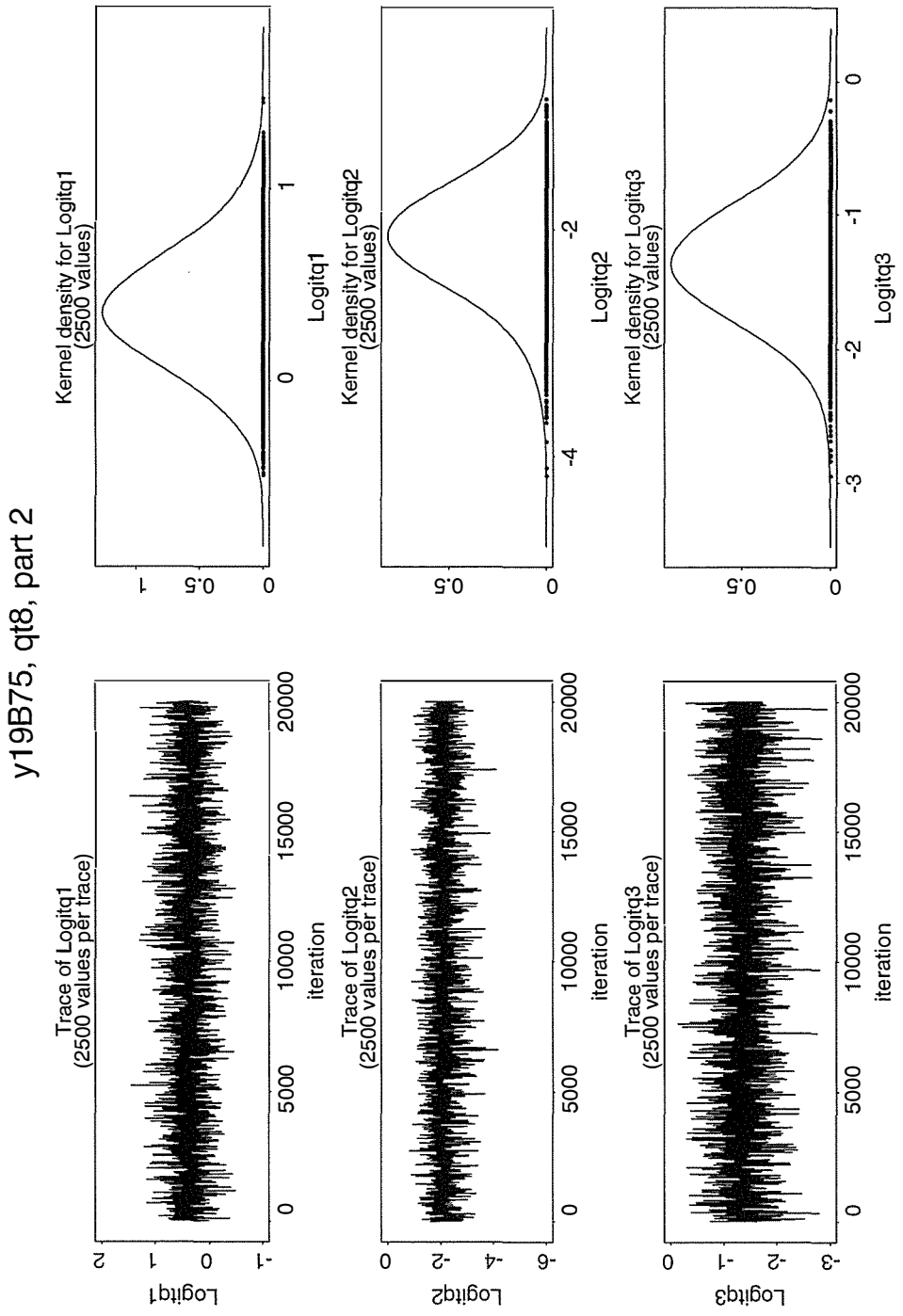


Figure D.11: Extension I, Experiment 2: MCMC Trace of the posterior distribution of effects of explanatory variables q_1, q_2, q_3

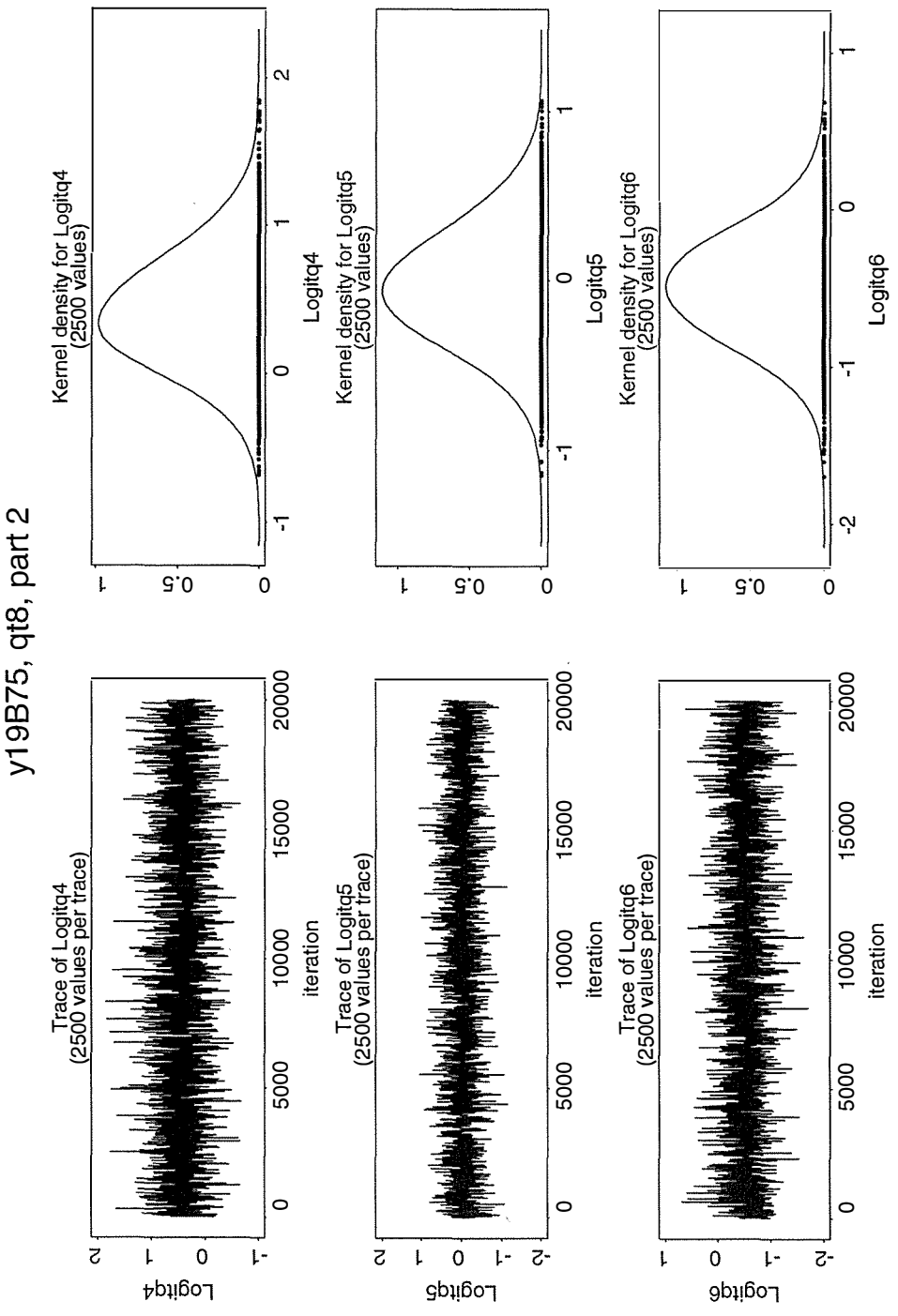


Figure D.12: Extension I, Experiment 2: MCMC Trace of the posterior distribution of effects of explanatory variables q_4, q_5, q_6

Table D.3: Extension I, Experiment 2: MCMC descriptive statistics of posterior distribution simulations of the effects of explanatory variables q_k.

Parameter	Sample Mean (Stdev[1])	50% Credible Interval
q_1	0.59 (0.067)	[0.54, 0.64]
q_2	0.11 (0.064)	[0.08, 0.14]
q_3	0.20 (0.080)	[0.16, 0.25]
q_4	0.60 (0.081)	[0.54, 0.66]
q_5	0.49 (0.080)	[0.44, 0.55]
q_6	0.38 (0.085)	[0.33, 0.44]

Table D.4: Extension I, Experiment 2: MCMC Convergence Diagnostics for effects of explanatory variables q_k

Parameter	Geweke Z	Raftery-Lewis	Heidelberger-Welch Tests CVM[2] Stationarity		Halfwidth	
α_1	0.222	1,3755	√	0.18	√	0.0042
α_2	-1.140	2,3729	√	0.07	√	0.0062
α_3	0.396	2,3760	√	0.05	√	0.0052
α_4	-0.992	2,3717	√	0.18	√	0.0047
α_5	-1.690	2.3732	√	0.10	√	0.0044
α_6	-1.430	2,3784	√	0.27	√	0.0043

Appendix E

MCMC diagnostics for *dingo* case study, Extension II

These diagnostics and summaries of posterior distributions are referred to in the text of Section 7.7.

Table E.1: Descriptive statistics for posterior distribution MCMC simulations of the effects of explanatory variables α and autoregressive time effect β.

Parameter	Sample Mean (Stdev)	SE (Naive TS) (Batch)	Lag 1 AC	Lag 1 Batch AC	IACT	50% Credible Interval
α_1	0.44 (0.344)	0.00243 0.00239 0.00261	0.0183	-0.0384	1.0645	[0.199, 0.659]
α_2	−2.27 (0.497)	0.00352 0.00337 0.00369	0.0113	-0.0009	0.9876	[-2.58, -1.92]
α_3	−1.47 (0.418)	0.00296 0.00282 0.00308	0.0244	-0.0460	1.0508	[-1.75, -1.19]
α_4	0.455 (0.427)	0.00302 0.00292 0.00314	0.0323	0.0090	1.0383	[0.158, 0.733]
α_5	−0.00638 (0.376)	0.00266 0.00266 0.00286	0.0385	0.0075	1.1328	[-0.265, 0.241]
α_6	−0.517 (0.367)	0.00260 0.00258 0.00266	0.0383	0.0634	1.0787	[-0.769, -0.270]
β	−0.421 (0.415)	0.00293 0.00278 0.00297	0.0106	-0.0054	1.0084	[-0.699, -0.141]

Table E.2: MCMC Convergence Diagnostics for effects of explanatory variables α and autoregressive time effect β

		Raftery-	Heidelberger-Welch Tests			
Parameter	Geweke Z	Lewis	Stationarity	CvM	Halfwidth	
α_1	0.05	2,3718	√	0.32	√	0.00469
α_2	0.29	1,3755	√	0.06	√	0.00660
α_3	1.49	2,3824	√	0.34	√	0.00552
α_4	0.87	2,3780	√	0.10	√	0.00573
α_5	-0.26	2,3842	√	0.11	×	0.00522
α_6	-0.70	1,3771	√	0.21	√	0.00505
β	-0.73	2,3802	√	0.11	√	0.00545

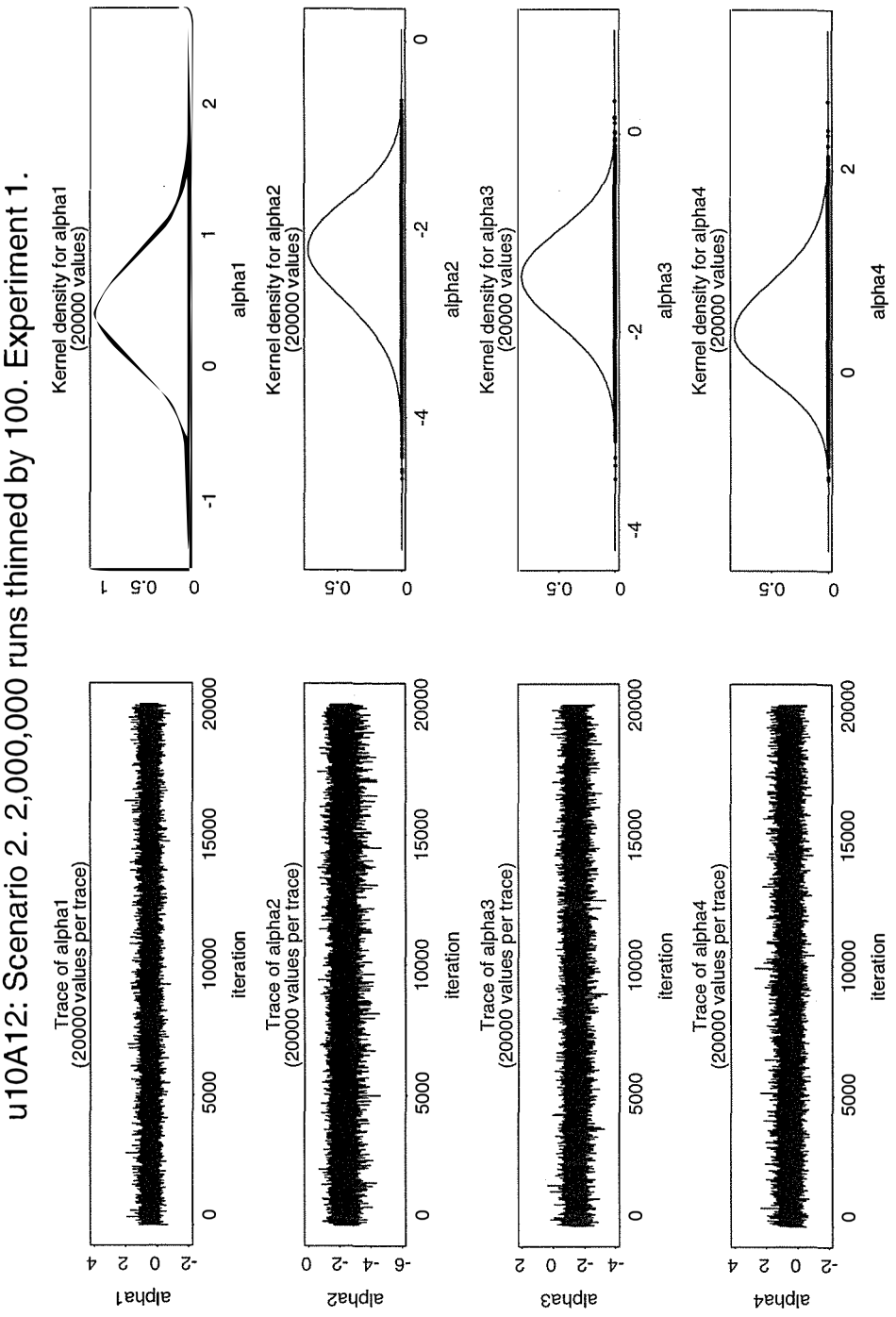


Figure E.1: Extension II, Experiment 1: MCMC Trace of the posterior distribution of parameters in linear predictor $\alpha_1 - -\alpha_4$

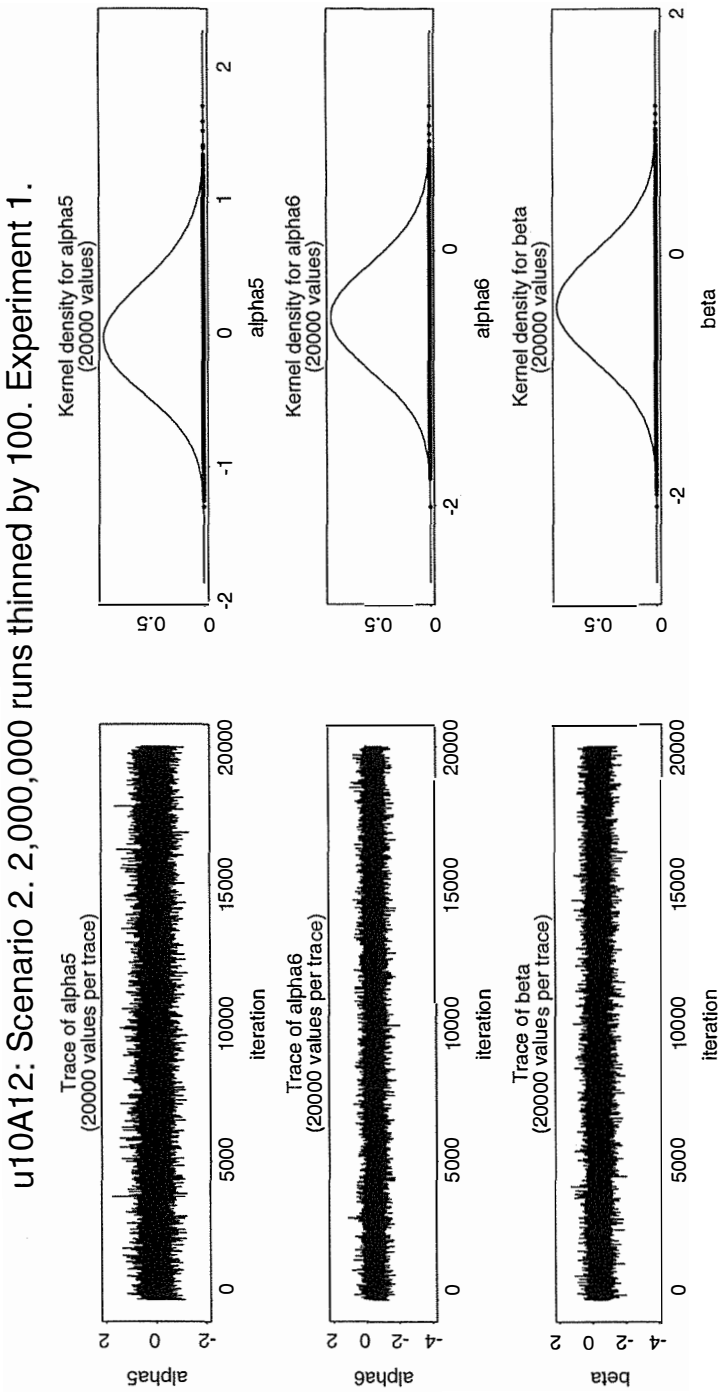


Figure E.2: Extension II, Experiment 1: MCMC Trace of the posterior distribution of parameters in linear predictor $\alpha_5, \alpha_6, \beta$

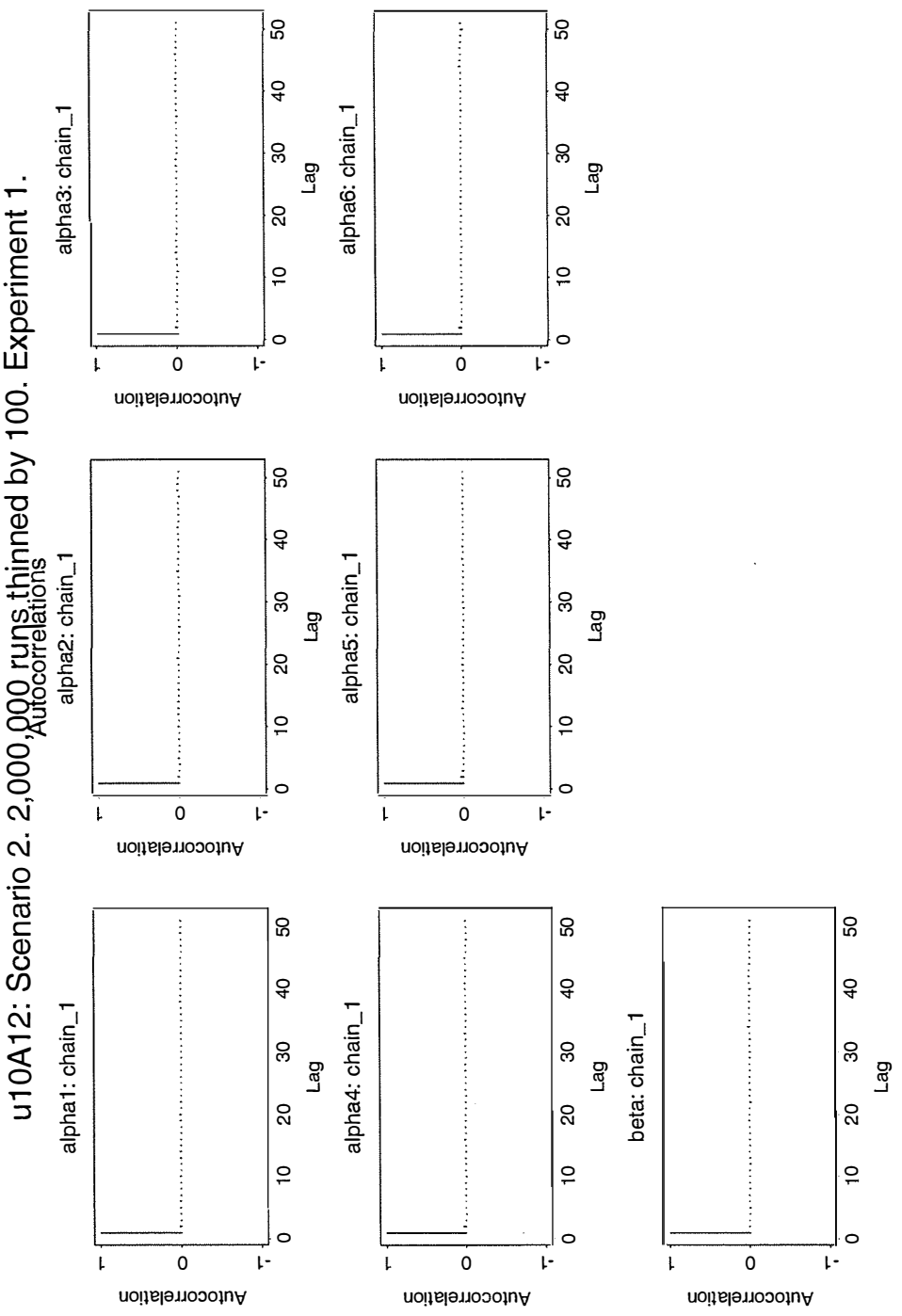


Figure E.3: Extension II, Experiment 1: MCMC Autocorrelation function for simulated values in posterior distribution of parameters in linear predictor α_k, β

Table E.3: Extension II, Experiment 1: Cross correlations between each α_k, β chain

Variable	α_1	α_2	α_3	α_4	α_5	α_6	β
α_1	1.000						
α_2	0.140	1.000					
α_3	0.181	0.099	1.000				
α_4	0.301	0.161	0.205	1.000			
α_5	0.285	0.146	0.193	0.325	1.000		
α_6	0.251	0.136	0.176	0.288	0.280	1.000	
β	-0.338	-0.106	-0.112	-0.325	-0.321	-0.232	1.000

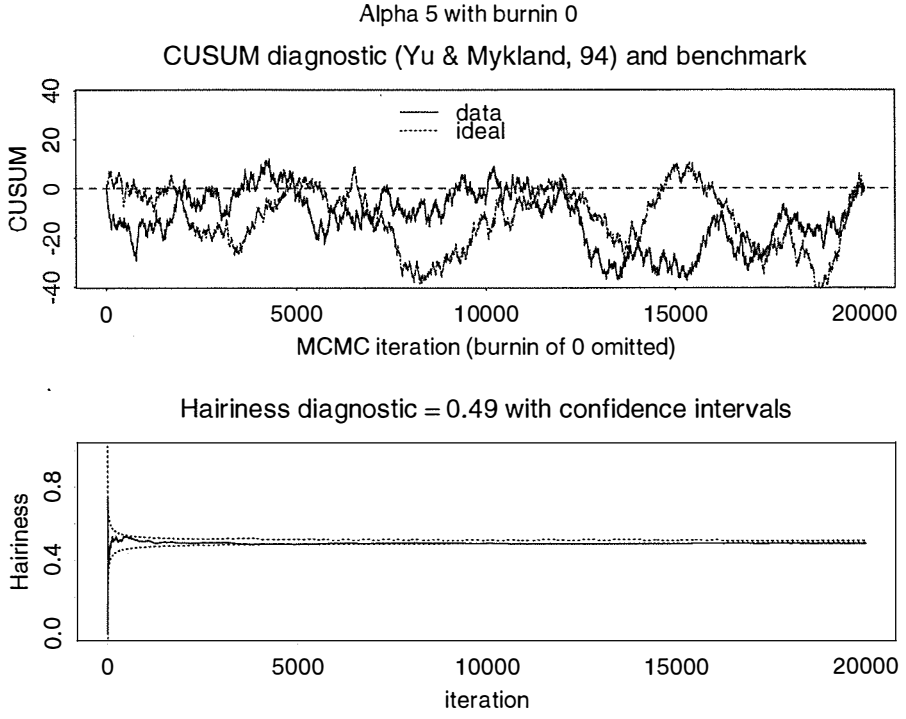


Figure E.4: Extension II, Experiment 1: CUSUM and hairiness diagnostic applied to simulation time series for parameter α_5

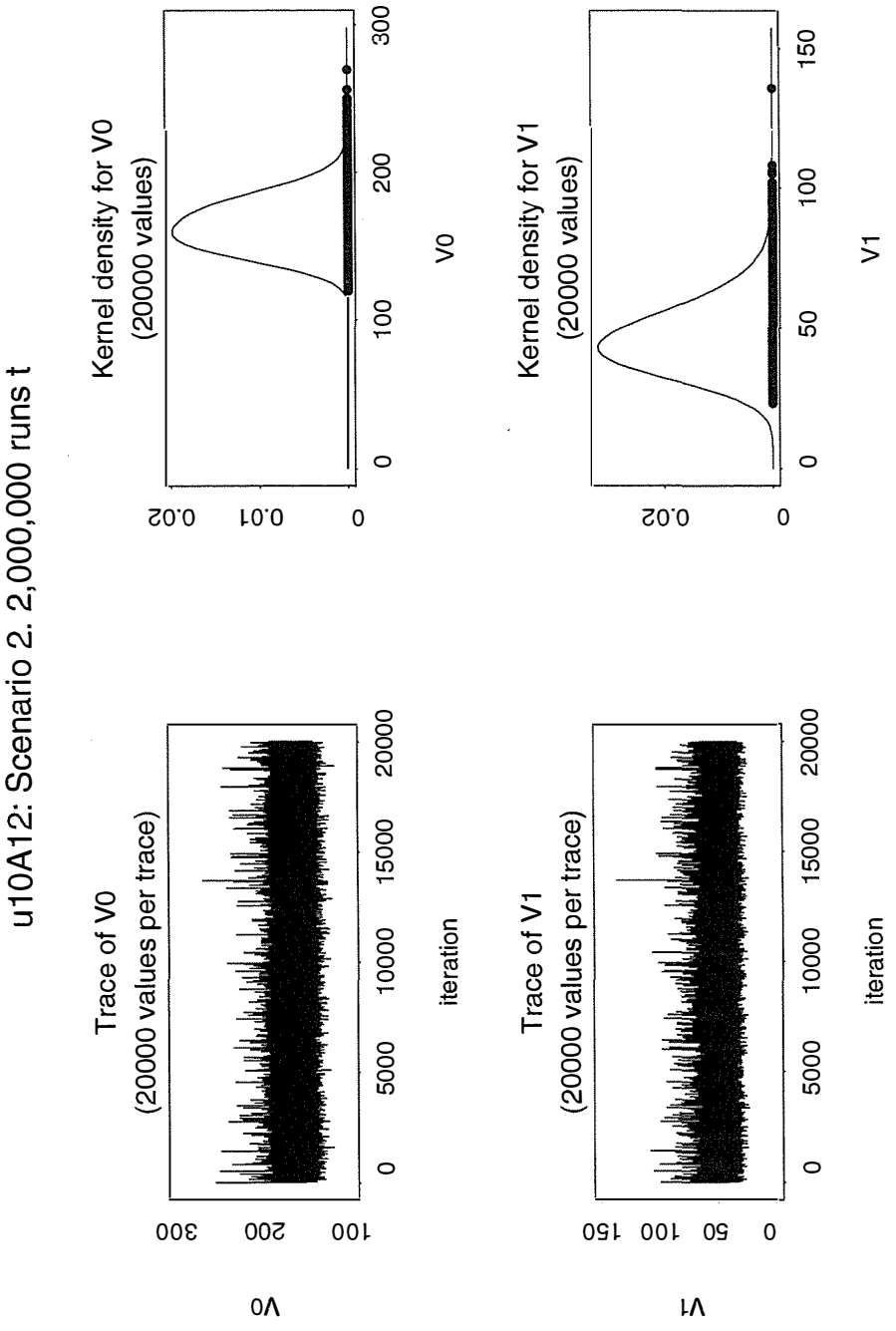


Figure E.5: Extension II, Experiment 1: MCMC Trace of the posterior distribution of dingo presence canonical statistics $V_0(z)$ and $V_1(z)$

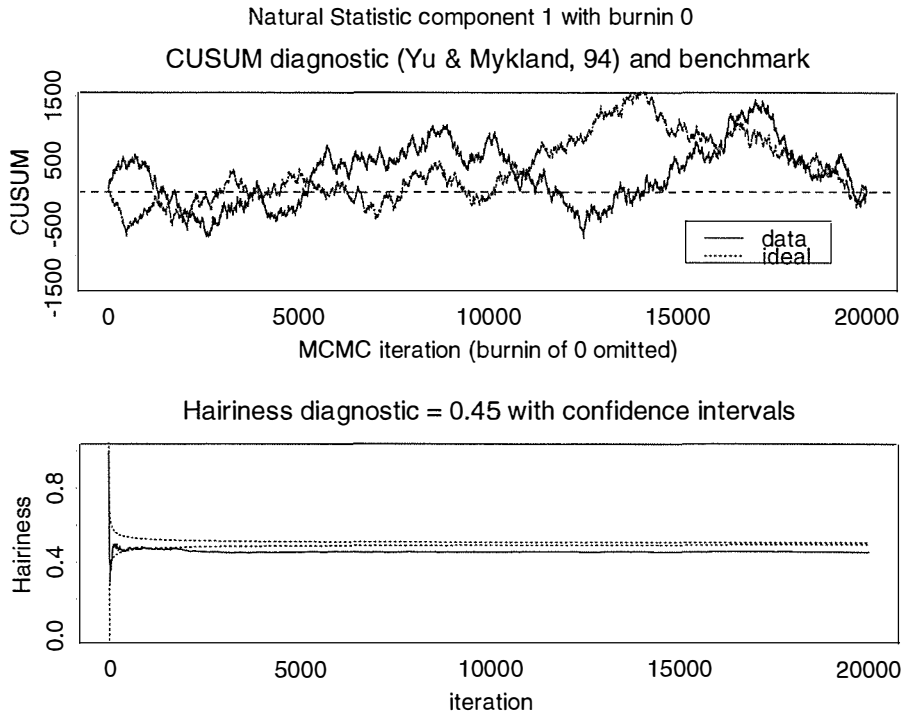


Figure E.6: Extension II, Experiment 1: CUSUM and hairiness diagnostic applied to simulation time series for parameter $V_1(z)$

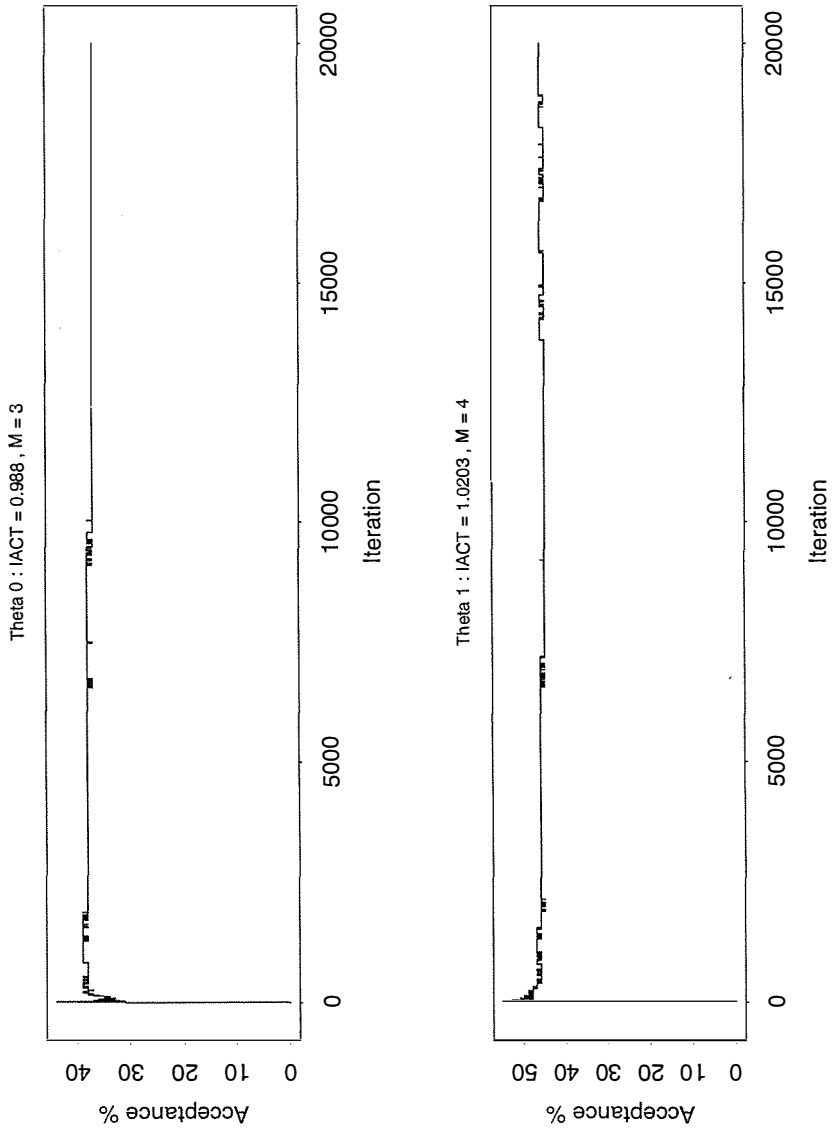


Figure E.7: Extension II, Experiment 1: Acceptance probability accumulated over simulation time for statistics $V_0(z)$ and $V_1(z)$

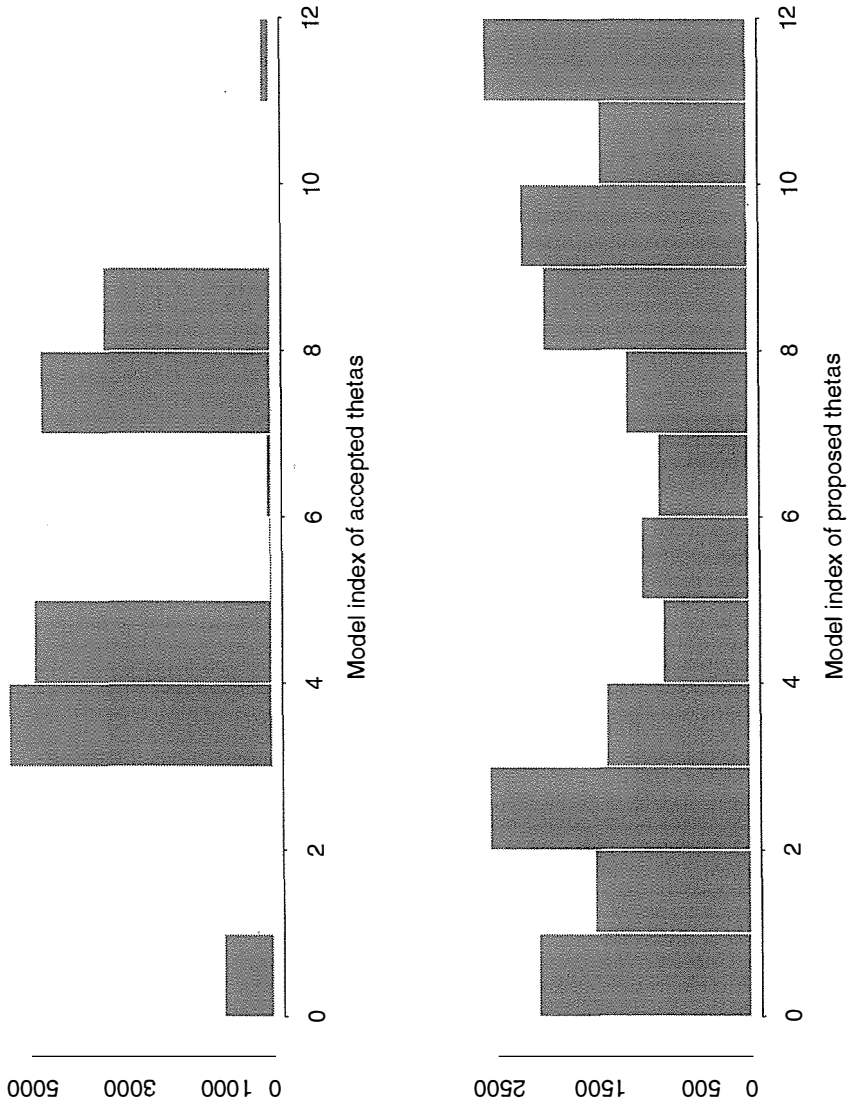


Figure E.8: Extension II, Experiment 1: Frequency of accepted and proposed model indexes corresponding to $\theta_{(m)}$.

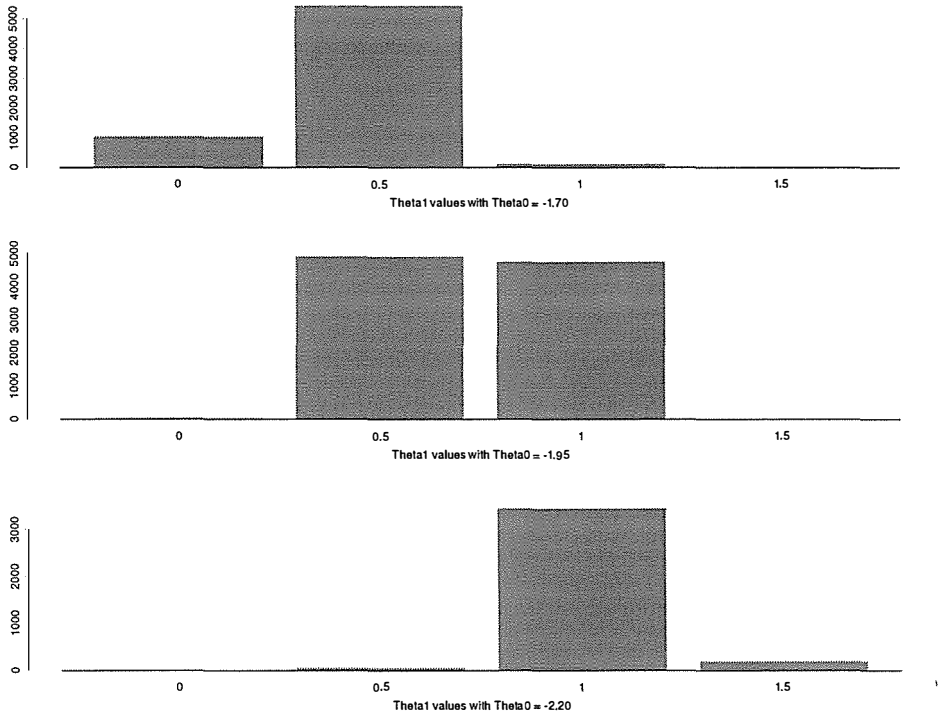


Figure E.9: Extension II, Experiment 1: Frequency of accepted and proposed θ components (θ_0, θ_1) corresponding to $\theta_{(m)}$.

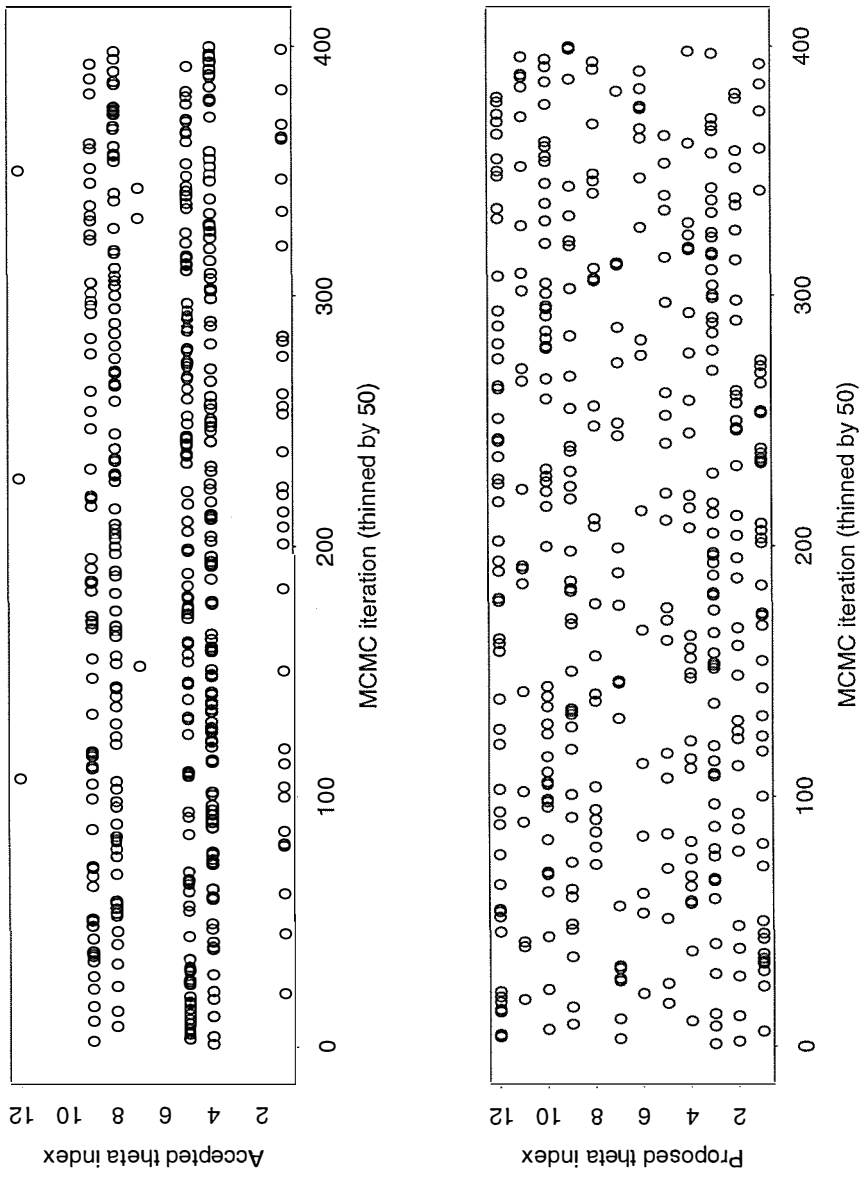


Figure E.10: Extension II, Experiment 1: Trace of accepted and proposed model indices m corresponding to $\theta_{(m)}$ over simulation time. To keep this plot visually accessible, m has only been plotted for every 50th iteration of 20,000 values (in addition to removal of burnin and thinning of original chain).

Table E.4: Extension II, Experiment 1: MCMC descriptive statistics of posterior distribution simulations of the effects of explanatory variables q_k.

Parameter	Sample Mean (Stdev)	SE (Naive) (TS) (Batch)	Lag 1 Batch AC	IACT	50% Credible Interval
q_1	0.551 (0.075)	0.00053 0.00050 0.00054	0.0003	1.00	[0.50, 0.60]
q_2	0.099 (0.041)	0.00029 0.00028 0.00030	0.0161	1.00	[0.07, 0.12]
q_3	0.185 (0.058)	0.00041 0.00039 0.00043	-0.0473	1.02	[0.14, 0.22]
q_4	0.554 (0.089)	0.00063 0.00060 0.00063	0.0338	0.98	[0.49, 0.61]
q_5	0.455 (0.079)	0.00056 0.00054 0.00057	0.0100	1.03	[0.40, 0.51]
q_6	0.347 (0.074)	0.00052 0.00051 0.00053	0.0660	1.00	[0.30, 0.39]

Table E.5: Extension II, Experiment 1: MCMC Convergence Diagnostics for effects of explanatory variables q_k

Parameter	Geweke Z	Raftery-Lewis	Heidelberger-Welch Tests Stationarity	CVM		Halfwidth
q_1	-0.28	2,3749	√	0.21	√	0.0010
q_2	0.13	1,3755	√	0.09	√	0.0005
q_3	1.35	2,3834	√	0.37	√	0.0008
q_4	0.46	2,3673	√	0.08	√	0.0012
q_5	-0.53	2,3842	√	0.10	√	0.0011
q_6	-1.01	1,3755	√	0.20	√	0.0010

Table E.6: Extension II, Experiment 1: MCMC diagnostics of posterior distribution simulations of the natural statistics for presence/absence $V(z)$.

Parameter	Sample Mean (Stdev)	SE (Naive) (TS) (Batch)	Lag 1 AC	Lag 1 Batch AC	IACT	50% Credible Interval
V_0	163 (16.4)	0.116 0.119 0.130	0.1090	-0.0039	1.03	[151, 175]
V_1	46.1 (10.8)	0.0765 0.0815 0.0851	0.0996	0.021	1.23	[38,53]

Table E.7: Extension II, Experiment 1: MCMC Convergence Diagnostics for presence/absence natural statistics $V(x)$

		Raftery-	Heidelberger-Welch Tests			
Parameter	Geweke Z	Lewis	Stationarity	CVM	Halfwidth	
$V_0(x)$	0.10	2,3843	$\surd$	0.37	$\surd$	0.234
$V_1(x)$	-0.66	1,4620	$\surd$	0.10	$\surd$	0.160

Bibliography

Abend, K., Harley, T. J. & Kanal, L. N. (1965), 'Classification of binary random patterns', *IEEE Trans. Inform. Theory* **11**, 533–544.

Agresti, A. (1990), *Categorical data analysis*, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons.

Aitchison, J. (1955), 'On the distribution of a positive random variable having a discrete probability mass at the origin', *J. Amer. Statist. Assoc.* **50**, 901–908.

Aitchison, J. & Ho, C. H. (1989), 'The multivariate Poisson-log normal distribution', *Biometrika* **476**, 643–53.

Aizenman, M. (1981), 'Proof of the triviality of φ_d^4 field theory and some mean-field features of Ising models for $d > 4$', *Phys. Rev. Lett.* **47**(1), 1–4.

Albert, J. & Chib, S. (1995), 'Bayesian residual analysis for binary response regression models', *Biometrika* **82**(4), 747–759.

Albert, P. S. & MacShane, L. M. (1995), 'A Generalized Estimating Equations approach for spatially correlated binary data. Applications to the analysis of neuroimaging data.', *Biometrics* **51**, 627–638.

Anderson, T. W. (1970), Estimation of covariance matrices which are linear combinations or whose inverses are linear combinations of given matrices, *in* 'Essays in Probability and Statistics', Univ. of North Carolina Press, Chapel Hill, N.C., pp. 1–24.

Anderssen, R. S., Latham, G. & Westcott, M. (1993), Statistical methodology for inverse problems, *in* S. Osaki & D. N. Pra Murthy, eds, 'Stochastic Models in Engineering, Technology and Management', World Scientific: Singapore, pp. 1–7.

Anselin, L. & Smirnov, O. (1998), *The SpaceStats Extension for ArcView.*

Aykroyd, R. G. & Green, P. J. (1991), 'Global and local priors, and the location of lesions using gamma-camera imagery', *Phil. Trans. R. Soc. Lond.* **337**(A), 323–342.

Baddeley, A. & Møller, J. (1989), 'Nearest-neighbour Markov point processes and random sets', *Int. Statist. Review* **57**(2), 89–121.

Baker, G. A. J. & Kawashima, N. (1996), 'Renormalized coupling constant in the Ising model', *J. Phys. A: Math. Gen.* **29**, 7183–7197.

Baker, Jr., G. A. (1961), 'Application of the Padé approximant method to the investigation of some magnetic properties of the Ising model', *Phys. Rev. (2)* **124**, 768–774.

Bartlett, M. S. (1971), 'Physical nearest-neighbour models and non-linear time-series', *J. Appl. Probability* **8**, 222–232.

Bartlett, M. S. (1978), 'Nearest neighbour models in the analysis of field experiments', *J. R. Statist. Soc. B* **40**(2), 147–174.

Baxter, R. J. (1982), *Exactly solved models in statistical mechanics*, Academic Press, London.

Becker, N. G. & Marshchener, I. C. (1993), 'A method for estimating the age-specific relative risk of HIV infection from AIDS incidence data.', *Biometrika* **80**, 165–78.

Becker, R. A., Chambers, J. M. & Wilks, A. R. (1988), *The new S language*, Wadsworth & Brooks/Cole, Pacific Grove, CA.

Bennett, C. H. (1976), 'Efficient estimation of free energy differences from Monte Carlo data', *J. Comput. Phys.* **22**, 245–268.

Bennett, J. H., ed. (1990), *Statistical methods, experimental design and statistical inference: collected works by R. A. Fisher.*, Oxford Science Publications.

Berkson, J. (1949), 'Minimum χ^2 and maximum likelihood solution in terms of a linear transform, with particular reference to bio-assay', *Journal of the American Statistical Association* **44**, 272–278.

Berkson, J. (1956), Estimation by least squares and by maximum likelihood, *in* 'Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I', University of California Press, Berkeley and Los Angeles, pp. 1–11.

Berkson, J. (1980), 'Minimum chi-square, not maximum likelihood', *The Annals of Statistics* **8**, 457–487.

Bernardo, J.-M. & Smith, A. F. M. (1994), *Bayesian theory*, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons Ltd., Chichester.

Bernardo, J. M., Berger, J. O., Dawid, A. P. & Smith, A. F. M., eds (1992), *Bayesian Statistics 4*, Oxford: University Press, Oxford.

Bernardo, J. M., Berger, J. O., Dawid, A. P. & Smith, A. F. M., eds (1999), *Bayesian Statistics 6*, Oxford University Press.

Besag, J. (1974), 'Spatial interaction and the statistical analysis of lattices systems (with discussion)', *J. R. Statist. Soc. B* **36**, 192–236.

Besag, J. (1975), 'Statistical analysis of non-lattice data', *The Statistician* **24**, 179–195.

Besag, J. (1986), 'On the statistical analysis of dirty pictures', *J. R. Statist. Soc. B* **48**(3), 259–302.

Besag, J. & Green, P. J. (1993), 'Spatial statistics and Bayesian computation', *J. R. Statist. Soc. B* **55**(1), 25–37.

Besag, J., Green, P., Higdon, D. & Mengersen, K. (1995), 'Bayesian computation and stochastic systems', *Statist. Sci.* **10**(1), 3–66. With comments and a reply by the authors.

Best, N., Cowles, M. K. & Vines, K. (1995), CODA Convergence Diagnostics and Output Analysis software for Gibbs sampling output version, 0.30, Lecture Notes for "An Introduction to Bayesian computation using BUGS", MRC Biostatistics Unit, Institute of Public Health, Cambridge.

Bethe, H. A. (1935), 'Statistical theory of superlattices', *Proc. R. Soc., London A* **150**, 552.

Binder, K. & Heermann, D. W. (1988), *Monte Carlo simulation in Statistical Physics: An Introduction*, number 80 *in* 'Springer series in Solid State Sciences', Springer-Verlag: Berlin, Hiedelberg.

Binder, K. & Heermann, D. W. (1992), *Monte Carlo simulation in statistical physics*, second edn, Springer-Verlag, Berlin. An introduction.

Binder, K. & Heermann, D. W. (1997), *Monte Carlo simulation in Statistical Physics: An Introduction*, number 80 *in* 'Springer series in Solid State Sciences', 3rd edn, Springer-Verlag: Berlin, Hiedelberg.

Binder, K., ed. (1986), *Monte Carlo methods in Statistical Physics*, Vol. 7 of *Topics Current Physics*, 2 edn, Springer: Berlin, Hiedelberg.

Bloemena, A. R. (1964), *Sampling from a graph*, Mathematisch Centrum, Amsterdam. Edited by W. R. van Zwet. Mathematical Centre Tracts, No. 2.

Böhning, D. (1999), *Computer-assisted analysis of mixtures and applications*, Chapman & Hall/CRC, Boca Raton, FL. Meta-analysis, disease mapping and others.

Box, G. E. P. & Jenkins, G. M. (1970), *Time Series Analysis, Forecasting and Control*, Holden-Day, San Francisco.

Box, G. E. P., Hunter, W. G. & Hunter, J. S. (1978), *Statistics for Experimenters: An introduction to Design, Data Analysis and Model building*, Wiley series in probability and mathematical statistics, John Wiley and sons.

Bragg, W. L. & Williams, E. J. (1934), *Proc. Soc. London Ser. A* **145**, 699.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984), *Classification and Regression trees*, The Wadsworth Statistics Probability Series, CRC Press, Monterey, California.

Breslow, N. E. (1981), 'Odds ratios when data is sparse', *Biometrika* **68**, 73–84.

Breslow, N. E. & Clayton, D. G. (1993), 'Approximate inference in generalized linear mixed models', *J. Amer. Statist. Assoc.* **88**, 9–25.

Breslow, N. E. & Day, N. E. (1987), *Statistical methods in cancer research*, International Agency for Research on Cancer, Lyon.

Brooks, S. (1997), Quantitative convergence diagnosis for MCMC via CUSUMS, Technical report, University of Bristol. WEB document: http://www.stats.bris.ac.uk/ maspb/mypapers/bro9bb.ps.

Brush, S. G. (1967), 'History of the Lenz-Ising model', *Reviews of Modern Physics* **39**(4), 883–893.

Buckland, S. T. & Elston, D. A. (1993), 'Empirical models for spatial distribution of wildlife', *Journal of Applied Ecology* **30**, 478–495.

Burley, D. M. (1972), Closed form approximations for lattice systems, *in* Domb & Green (1972*b*), pp. 329–373.

Carlin, B. P. & Chib, S. (1995), 'Bayesian model choice via Markov chain Monte Carlo methods', *J. R. Statist. Soc. B* **57**(3), 473–484.

Caughley, G., Short, J., Grigg, G. C. & Nix, H. (1987), 'Kangaroos and climate: An analysis of distribution', *Journal of Animal Ecology* **56**, 751–761.

Chakraborty, S., Pettitt, A. N., Boland, R. M., Low Choy, S., Cameron, D. F., Irwin, J. A. G. & Davis, R. D. (1993), Stylo host heterogeneity for anthracnose management, *in* 'Proceedings, XVII International Grassland Congress'.

Chakraborty, S., Pettitt, A. N., Cameron, D. F., Irwin, J. A. G. & Davis, R. D. (1991), 'Anthracnose development in pure and mixed stands of the pasture legume *stylosanthes scabra*', *Phytopathology* **81**(7), 788–793.

Chakraborty, S., Pettitt, A. N., Low Choy, S. & Boland, R. M. (1995), 'Spatial dependence in Anthracnose development in mixtures of *stylosanthes scabra*', *Phytopathology* **143**, 693–699.

Chen, C.-C. (1988), Markov Random Fields in Image Analysis, PhD thesis, Department of Computer Science.

Chen, M.-H. & Shao, Q. M. (1997), 'On Monte Carlo methods for estimating ratios of normalizing constants', *Ann. Statist.* **25**, 1563–1594.

Chiou, J.-M. & Müller, H.-G. (1998), 'Quasi-likelihood regression with unknown link and variance functions', *Journal of the American Statistical Association* **93**(444), 1376–1387.

Clark, L. & Pregibon, D. (1992), Tree based models, *in* C. J.M. & H. T.J., eds, 'Statistical Models in S', Wadworth & Brooks/Cole, chapter 9, pp. 377–419.

Cook, D. G. & Pocock, S. J. (1983), 'Multiple regression in geographic mortality studies with allowance for spatially correlated errors', *Biometrics* **39**, 361–371.

Corbett, L. K. (1995), *The Dingo in Australia and Asia*, University of New South Wales Press Ltd., Sydney.

Cowles, M. K. & Carlin, B. P. (1996), 'Markov Chain Monte Chain convergence diagnostics: A comparative review', *J. Amer. Statist. Assoc.* **91**(434), 833–904.

Cox, D. R. (1970), *The analysis of binary data*, Chapman and Hall: London.

Cox, D. R. & Hinkley, D. V. (1974), *Theoretical Statistics*, Chapman & Hall: London.

Cressie, N. A. C. (1985), 'Fitting variogram models by weighted least squares.', *Journal of the International Association for Mathematical Geology* **17**, 563–586.

Cressie, N. A. C. (1993), *Statistics for Spatial Data*, John Wiley & Sons.

Del Grosso, G. (1974), 'On the local central limit theorem for Gibbs processes', *Comm. Math. Phys.* **37**, 141–160.

Dellaportas, P. (1995), 'Random variate transformations in the Gibbs sampler: Issues of efficiency and convergence', *Statistics and Computing* **5**, 133–140.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *J. R. Statist. Soc. B* **39**, 1–38.

Denham, R. & Mengersen, K. (1999), Bayesian spatial logistic regression for remote sensing, unpublished.

Derin, H. & Cole, W. S. (1986), 'Segmentation of textured images using Gibbs Random Fields', *Computer Vision, Graphics and Image Processing* **35**, 72–98.

Derin, H. & Elliott, H. (1987), 'Modeling and segmentation of noisy and textured images using Gibbs random fields', *IEEE Transactions on Pattern Anal. & Machine Intell.* **6**(6), 39–55.

Diggle, P. (1983), *Statistical analysis of spatial point patterns*, London: Academic Press.

Diggle, P. & Gratton, R. J. (1984), 'Monte Carlo methods of inference for implicit statistical models', *J. R. Statist. Soc. B* **46**(2), 193–227.

Diggle, P. J., Fiksel, T., Grabarnik, P., Ogata, Y., Stoyan, D. & Tanemura, M. (1994), 'On parameter estimation for pairwise interaction point processes', *International Statistical Review* **62**(1), 99–117.

Diggle, P. J., Liang, K.-L. & Zeger, S. L. (1996), 'Analysis of longitudinal data'. (2nd edition).

Diggle, P., Tawn, J. & Moyeed, R. (1998), 'Model-based geostatistics', *Appl. Statist.* **47**, 299–350.

Dobrushin, R. L. (1968), 'The description of a random field by means of conditional probabilities and conditions of its regularity', *Theory of Probability and its Applications* **13**, 197–224.

Domb, C. & Green, M., eds (1972*a*), *Phase Transitions and Critical Phenomena*, Vol. 1. Exact Results, Academic Press, London.

Domb, C. & Green, M., eds (1972*b*), *Phase Transitions and Critical Phenomena*, Vol. 2., Academic Press, London.

Domb, C. & Green, M., eds (1974), *Phase Transitions and Critical Phenomena*, Vol. 3. Series expansions for lattice models, Academic Press, London.

Domb, C. & Green, M., eds (1976), *Phase Transitions and Critical Phenomena*, Vol. 6, Academic Press, London.

Domb, C. & Green, M. S. (1972*c*), Preface to volume 1., *in* Domb & Green (1972*a*), pp. ix–xii.

Dubes, R. C. & Jain, A. K. (1989), 'Random field models in image analysis', *Journal of Applied Statistics* **16**(2), 131–163.

Dubois, G. (2001), *AI-Geostats*, WWW, http://curie.ei.jrc.it/ai-geostats.htm.

Efron, B. (1979), 'Bootstrap methods: another look at the Jacknife', *Annals Stat.* **7**(1), 1–26.

Environmental Systems Research Institute (1996), *Using ArcView GIS*.

Environmental Systems Research Institute (1997), *Understanding GIS; The ARC/INFO Method*, version 7.1 edn, UK.

ERDAS (1998), *Imagine Reference Manual*.

Fan, J. & Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, number 66 *in* 'Monographs on Statistics and Applied Probability', Chapman & Hall: London.

Flyvbjerg, H. & Petersen, H. G. (1989), 'Error estimates on averages of correlated data', *J. Chem. Phys.* **91**(1), 461–466.

Föllmer, H. (1982), 'A covariance estimate for Gibbs measures', *Journal of Functional Analysis* **46**, 387–395.

Forsythe, G. E., Malcolm, M. A. & Moler, C. B. (1977), *Computer Methods for Mathematical Computations*, Prentice-Hall series in Automatic Computation, Prentice-Hall: New Jersey.

Gaunt, D. S. & Guttmann, A. J. (1974), Asymptotic analysis of coefficients, *in* Domb & Green (1974), pp. 181–241.

Gelfand, A. & Dey, D. (1994), 'Bayesian model choice: Asymptotics and exact calculations', *J. R. Statist. Soc. B* **56**(3), 501–514.

Gelfand, A. E. & Smith, A. F. M. (1990), 'Sampling based approaches to calculating marginal densities', *J. Amer. Statist. Assoc.* **85**, 398–409.

Gelfand, A. E., Sahu, S. K. & Carlin, B. P. (1994), Efficient parameterizations for normal linear mixed models, *in* 'Bayesian Statistics 5', pp. 165–180.

Gelfand, A. E., Sahu, S. K. & Carlin, B. P. (1995), Efficient parameterizations for generalized linear mixed models, Unpublished manuscript. Obtained from World Wide Web MCMC Preprints.

Gelman, A. & Meng, X.-L. (1994), Path sampling for computing normalizing constants: identities and theory., Technical Report 376, Department of Statistics, University of Chicago.

Gelman, A. & Meng, X.-L. (1996), Simulating normalizing constants: from importance sampling to bridge sampling to path sampling., Technical Report 440, Department of Statistics, University of Chicago.

Gelman, A. & Meng, X.-L. (1998), 'Simulating normalizing constants: from importance sampling to bridge sampling to path sampling', *Statistical Science* **13**(2), 163–182.

Gelman, A. & Rubin, D. B. (1992), 'Inference from iterative simulation using multiple sequences', *Statistical Science* **7**(4), 457–511.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (1995), *Bayesian Data Analysis*, Texts in Statistical Science, Chapman and Hall: Great Britain.

Geman, S. & Geman, D. (1984), 'Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**(6), 721–741.

George, E. L. & McCulloch, R. E. (1993), 'Variable selection via Gibbs sampling', *J. Amer. Statist. Assoc.* Theory and Methods.

Georgi, H. O. Haggstrom, O. (1996), 'Phase transition in continuum Potts models', *Commun. Math. Phys.* **181**, 507–528.

Geweke, J. (1992), Evaluating the accuracy of sampling-based approaches to calculating posterior moments, *in* Bernardo, Berger, Dawid & Smith (1992).

Geyer, C. J. (1994), Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo, Technical Report 568, School of Statistics, University of Minnesota.

Geyer, C. J. (1996), Estimation and optimization of functions, *in* Gilks et al. (1996), pp. 241–258.

Geyer, C. J. & Thompson, E. A. (1992), 'Constrained Monte Carlo Maximum Likelihood for dependent data', *J. R. Statist. Soc. B* **54**(3), 657–699.

Geyer, C. J. & Thompson, E. A. (1995), 'Annealing Markov Chain Monte Carlo with applications to ancestral inference', *J. Amer. Statist. Assoc.* **90**(431), 909–920.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J., eds (1996), *Markov Chain Monte Carlo in Practice*, Chapman & Hall, London.

Glauber, R. J. (1963), 'Time-dependent statistics of the Ising model', *J. Mathematical Phys.* **4**, 294–307.

Glötzl, E. & Rauchenschwandtner, B. (1981), On the statistics of Gibbsian processes, *in* P. Revesz, L. Schmetterer & V. Zolotarev, eds, 'The first Pannonian symposium on Mathematical Statistics', Vol. 8 of *Lecture Notes in Statistics*, Springer-Verlag.

Graybill, F. A. (1983), *Matrices with application in statistics, 2nd edition*, Wadsworth: Belmont, California.

Green, P. J. (1984), 'Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives (with discussion)', *J. R. Statist. Soc. B* **46**(2), 149–192.

Green, P. J. (1990), 'Bayesian reconstructions from emission tomography data using a modified EM algorithm', *IEEE Transactions on Medical Imaging* **9**(1), 84–93.

Green, P. J. (1995), 'Reversible jump Markov chain Monte Carlo computation and Bayesian model determination', *Biometrika* **82**(4), 711–32. via browsing.

Green, P. J. & Han, X.-l. (1990), *Metropolis methods, Guassian proposals and antithetic variables.*, number 74 *in* 'Lecture Notes in Statist., ed. P. Barone, A. Frigessi and M. Piccioni', Springer, Rome, pp. 142–164.

Green, P. J. & Murdoch, D. J. (1999), Exact sampling for Bayesian inference: Towards general purpose algorithms, *in* Bernardo, Berger, Dawid & Smith (1999), pp. 301–322.

Green, P., Jennison, C. & Seheult, A. (1985), 'Analysis of field experiments by least squares smoothing', *J. R. Statist. Soc. B* **47**(2), 299–315.

Griffiths, R. B. (1964), 'Peierl's proof of spontaneous magnetization in two dimensional Ising ferromagnet', *Phys. Rev. A* **136**, 437–438.

Grimmett, G. (1973), 'A theorem about random fields', *Bulletin of the London Mathematical Society* **5**, 81–84.

Guénault, T. (1995), *Statistical Physics*, 2nd edn, Chapman and Hall, London.

Häggström, O. & Nelander, K. (1997*a*), 'Exact sampling from anti-monotone systems', *Statistica Neerlandica* **0**, 0.

Häggström, O. & Nelander, K. (1997*b*), 'On exact simulation of Markov random fields using coupling from the past', *Scandinavia Journal of Statistics* **0**, 0.

Häggström, O., van Lieshout, M. N. M. & Møller, J. (1999), 'Characterization results and Markov chain Monte Carlo algorithms including exact simulation for some spatial point processes', *Bernoulli* **0**, 0.

Hall, P. (1985), 'Resampling a coverage process', *Stoch. Proc. Applic.* **20**, 231–46.

Hall, P., Horowitz, J. L. & Jing, B.-Y. (1995), 'On blocking rules for the bootstrap with dependent data', *Biometrika* **87**, 561–74.

Hammersley, J. & Mazzarino, G. (1983), Markov fields, correlated percolation, and the Ising model, *in* B. D. Hughes & B. V. Ninham, eds, 'The Mathematics and Physics of disordered media', Vol. 1035 of *Lecture Notes in Mathematics*, Springer-Verlag.

Hammersley, J. M. & Handscomb, D. C. (1964), *Monte Carlo Methods*, London: Methuen.

Hastings, W. K. (1970), 'Monte Carlo sampling methods using Markov Chains and their applications', *Biometrika* **57**(1), 97–109.

Hay, J. L. (1999), Statistical Modelling for non-Gaussian Time Series Data with Explanatory Variables, PhD thesis, School of Mathematical Sciences.

Heikkinen, J. & Högmander, H. (1994), 'Fully Bayesian approach to image restoration with an application in biogeography', *Appl. Stat.* **43**(4), 569–582.

Hiedelberger, P. & Welch, P. (1983), 'Simulation run length control in the presence of an initial transient', *Operations Research* **31**, 1109–1144.

Hill, M. O. (1991), 'Patterns of species distribution in Britain elucidated by canonical correspondence analysis', *Journal of Biogeography* **18**, 247–255.

Hills, S. E. & Smith, A. F. M. (1992), Parameterization issues in Bayesian inference, *in* Bernardo et al. (1992), pp. 227–246.

Hjort, N. L. & Omre, H. (1994), 'Topics in spatial statistics', *Scand. J. Statist.* **21**, 289–357.

Hoëting, J. A., Van Caster, M. & Bowden, D. (1997), An improved model for spatially correlated binary responses, WWW document, Colorado State University Statistics Depatment.

Högmander, H. & Møller, J. (1995), 'Estimating distribution maps from atlas data using methods of statistical image analysis', *Biometrics* **51**, 393–404.

Huang, K. (1987), *Statistical Mechanics*, 2nd edn, John Wiley & Sons, New York.

Huffer, F. W. & Wu, H. (1998), 'Markov chain Monte Carlo for autologistic regression models with application to the distribution of plant species', *biometrics* **54**, 509–524.

Hurst, C. A. & Green, H. S. (1960), 'New solution of the Ising problem for a rectangular lattice', *J. Chem. Phys.* **33**, 1059–1062.

Ickstadt, K. & Wolpert, R. (1999), Spatial regression for marked point processes, *in* Bernardo et al. (1999), pp. 323–342.

Ising, E. (1925), 'Bietrag zur theorie des ferromagnetismus', *Zeitschrift für Physik* **31**, 253–258.

Johnson, N. L. & Kotz, S. (1969), *Distributions in statistics: discrete distributions*, Houghton Mifflin, Boston.

Kac, M. (1964), 'The work of T. H. Berlin in statistical mechanics—a personal reminiscence', *Phys. Today.*

Kac, M. & Ward, J. C. (1952), 'A combinatorial solution of the two-dimensional Ising model', *Phys. Rev.*

Kass, R. E. & Slate, E. H. (1992), Reparameterization and diagnostics of posterior non-normality, *in* Bernardo et al. (1992), pp. 289–305.

Kastelyn, P. W. (1963), *J. Math. Phys.* **4**, 287.

Kawasaki, K. (1972), Kinetics of Ising model, *in* Domb & Green (1972*b*), pp. 443–498.

Kennedy, W. J. & Gentle, J. E. (1980), *Statistical Computing*, Marcel Dekker, New York.

Kindermann, R. & Snell, J. L. (1980), *Markov random fields and their applications*, American Mathematical Society, Providence, R.I.

Kramers, H. A. & Wannier, G. H. (1941), 'Statistics of the two-dimensional ferromagnet. I–II.', *Physical Review* **60**, 252–278.

Kiinsch, H. (1983), 'Asymptotically unbiased inference for Ising models', *Advances in Applied Probability* **15**, 887–888.

Lambert, D. (1992), 'Zero-inflated Poisson regression, with an application to defects in manufacturing', *Technometrics* **34**(1), 1–14.

Landau, D. P. & Tang, S. (1988), *J. de Phys.* **49**, C8–1525.

Lavine, M. (1999), Another look at conditionally Gaussian Markov random fields, *in* Bernardo et al. (1999), pp. 371–388.

Lebowitz, J. L. (1974), *Commun. Math. Phys.* **35**, 87.

Lee, P. M. (1989), *Bayesian Statistics: an Introduction*, Oxford University Press: New York.

Lenz, W. (1920), *Physik Z.* **21**, 613.

Lewis, S. M. & Raftery, A. E. (1997), 'Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator', *J. Amer. Statist. Assoc.* **92**(438), 648–655. previously unpublished 1994, preprint from World Wide Web.

Liang, K. Y. & Zeger, K. S. (1986), 'Longitudinal data analysis using generalized linear models', *Biometrika* **73**, 13–22.

Lindley, D. V. & Smith, A. F. M. (1972), 'Bayes estimates for the linear model', *J. Roy. Statist. Soc. Ser. B* **34**, 1–41. With discussions by J. A. Nelder, V. D. Barnett, C. A. B. Smith, T. Leonard, M. R. Novick, D. R. Cox, R. L. Plackett, P. Sprent, J. B. Copas, D. V. Hinkley, E. F. Harding, A. P. Dawid, C. Chatfield, S. E. Fienberg, B. M. Hill, R. Thompson, B. de Finetti, and O. Kempthorne.

Louis, T. A. (1982), 'Finding the observed information matrix when using the EM algorithm', *J. R. Statist. Soc. B* **44**(2), 226–233.

Low Choy, S. & Pettitt, A. N. (1992), Analysis of effect of different genotypes of *stylosanthes scabra* in promotion/reduction of resistance to anthracnose. Focus on plants of genotype 55860., Research Report for Dr S. Chakraborty, Department of Botany, University of Queensland, Statistical Consulting unit, Queensland University of Technology.

Low Choy, S. J. & Pettitt, A. N. (1997), 'Hierarchical Bayesian models of underlying spatial dependence for binary lattice data with missing observations. Motivation: analysis of data from a field trial investigating effectiveness of chemical attractants on Australian dingoes.', Poster presented and awarded first prize at Bayes 6; judged by A. Gelfand. A web-friendly version of the contents of the poster is available at http://www.math.fsc.qut.edu.au/poster/contents.htm.

MacDonald, I. L. & Zucchini, W. (1997), *Hidden Markov and other models for discrete-valued time series*, Chapman and Hall, London.

Madigan, D. & York, J. (1995), 'Bayesian Graphical Models for Discrete Data', *Int. Statist. Rev.* **63**(2), 215–32.

MAP INFO Corporation (1997), *MAP INFO User's guide, Version 5.0.*

Matherson, G. (1963), 'Principles of geostatistics', *Economic Geology* **58**, 1246–1266.

McCullagh, P. & Nelder, J. A. (1989), *Generalized Linear Models*, Chapman and Hall: London.

McCullagh, P. & Nelder, J. A. (1993), *Generalized Linear Models*, 2nd edn, Chapman and Hall: London.

Meng, X. L. & Wong, W. H. (1996), 'Simulating ratios of normalizing constants via a simple identity: a theoretical exploration.', *Statistica Sinica* **6**, 831–860.

Mengersen, K. L., Roberts, C. P. & Guihenneuc-Jouyaux, C. (1999), MCMC convergence diagnostics: A reviewww, *in* Bernardo et al. (1999), pp. 415–440.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953), 'Equations of state calculations by fast computing machines', *J. Chem. Phys.* **21**, 1087–1092.

Meyn, S. P. & Tweedie, R. L. (1993), *Markov chains and stochastic stability*, Springer-Verlag London Ltd., London.

Mitchell, J. (1988), Dingoes ...., Technical report, Queensland Lands Department.

Moran, P. A. P. (1947), 'Random associations on a lattice', *Proceedings of the Cambridge Philosophical Society* **43**, 321–328.

Moussouris, J. (1974), 'Gibbs and Markov random systems with constraints', *J. Statist. Phys.* **10**, 11–33.

Mullahy, J. (1986), 'Specification and testing of some modified count data models', *J. Econometrics* **33**(3), 341–365.

Müller-Krumbhaar, H. & Binder, K. (1973), *J. Stat. Phys.*

Muller, P. (1994), A generic approach to posterior integration and Gibbs sampling., unpublished manuscript.

Mykland, P., Tierney, L. & Yu, B. (1992), Regeneration in Markov chain samplers, Technical Report 585, University of Minnesota. Technical Report.

Newton, M. A. & Raftery, A. E. (1994), 'Approximate Bayesian inference with the weighted likelihood bootstrap', *J. R. Statist. Soc. B* **56**(1), 3–48.

Niemeijer, T. & van Leeuwen, J. M. J. (1976), Renormalization theory for Ising-like spin systems, *in* Domb & Green (1976), pp. 425–502.

Nychka, D. (1990), 'Some properties of adding a smoothing step to the EM algorithm', *Statist. Probab. Lett.* **9**, 187–193.

Ogata, Y. & Tanemura, M. (1981), 'Estimation of interaction potentials of spatial point patterns through the maximum likelihood procedure', *Annals of the Institute of Statistical Mathematics* **33**, 315–338.

Ogata, Y. & Tanemura, M. (1984), 'Likelihood analysis of spatial point patterns', *Journal of the Royal Statistical Society* **Ser. B, 46**(3), 496–518.

Onsager, L. (1944), 'Crystal statistics, I. A two-dimensional model with an order-disorder transition', *Phys. Rev.*. **65**, 117–149.

Osborne, P. E. & Tigar, B. J. (1992), 'Interpreting bird atlas data using logistic models: An example from Lesotho, Southern Africa', *Journal of Applied Ecology* **29**, 55–62.

Patterson, H. D. & Thompson, R. (1971), 'Recovery of inter-block information when block sizes are unequal', *Biometrika* **58**, 545–554.

Payne, R. W., Lane, P. W., Ainsley, A. E., Bicknell, K. E., Digby, P. G. N., Harding, S. A., Leech, P. K., Simpson, H. R., Todd, A. D., Verrier, P. J. & White, R. P. (1987), *Genstat 5 Reference Manual*, Oxford.

Pearcall, J., ed. (1998), *The New Oxford dictionary of English*, Clarendon Press: Oxford.

Peierls, R. E. (1936), 'On Ising's ferromagnet model', *Proc. Camb. Phil. Soc.* **32**, 477–481.

Pendergast, J. F., Gange, S. J., Newton, M. A., Lindstrom, M. J., Palta, M. & Fisher, M. A. (1996), 'A survey of methods for analyzing clustered binary response data', *International Statistical Review* **64**, 89–118.

Pettitt, A. N. & Low Choy, S. (1999), 'Bivariate binary data with missing values: Analysis of a field experiment to investigate chemical attractants of wild dogs', *J. Agric. Biol. and Environ. Statist.* **4**(1), 57–76.

Pickard, D. (1976), 'Asymptotic inference for an Ising lattice', *J. Appl. Prob.* **1976**, 486–497.

Pickard, D. (1977*a*), 'Asymptotic inference for an Ising lattice II.', *Adv. Appl. Prob.* **9**, 476–501.

Pickard, D. (1977*b*), 'A curious binary lattice process', *Journal of Applied Probability* **16**, 12–24.

Pickard, D. (1979), 'Asymptotic inference for an Ising lattice III. Non-zero field and ferromagnetic states', *J. Appl. Prob.* **16**, 12–24.

Pickard, D. (1982), 'Inference for general Ising models', *Journal of Applied Probability (Special Volume)* **19A**, 345–357.

Pickard, D. K. (1987), 'Inference for discrete Markov fields: the simplest non-trivial case', *J. Amer. Statist. Assoc.* **82**(397), 90–96.

Possolo, A. (1986*a*), Estimation of binary Markov random fields, Technical Report 77, Department of Statistics, University of Washington. Seattle, Washington.

Possolo, A. (1986*b*), Subsampling a random field, Technical Report 78, Department of Statistics, University of Washington.

Preisler, H. K. (1993), 'Modelling spatial patterns of trees attacked by bark-beetles', *Appl. Statist.* **43**(3), 510–514.

Priestley, M. B. (1981), *Spectral Analysis and Time Series*, Academic Press, London.

Propp, J. G. & Wilson, D. B. (1996), 'Exact sampling with coupled Markov chains and applications to statistical mechanics', *Random structures and Algorithms* **9**, 223–252.

Qian, W. & Titterington, D. (1991), 'Multidimensional Markov Chain Models for Image Textures', *J. R. Statist. Soc. B* **53**, 661–674.

Raftery, A. E. & Lewis, S. M. (1992), How many interations in the Gibbs sampler?, *in* Bernardo et al. (1992), pp. 765–776.

Rao, C. R. (1971*a*), 'Estimation of variance and covariance components–MINQUE theory', *J. Multiv. Anal.* **1**, 257–275.

Rao, C. R. (1971*b*), 'Minimum variance quadratic unbiased estimation of variance components', *J. Multiv. Anal.* **1**, 445–456.

Reynolds, K. M., Madden, L. V. & Ellis, M. A. (1988), 'Spatio-temporal analysis of epidemic development of leather rot of strawberry', *Phytopathology* **78**(2), 246–252.

Ripley, B. D. (1981), *Spatial statistics*, Wiley series in probability and mathematical statistics, John Wiley & Sons, Inc.

Ripley, B. D. (1988), *Statistical inference for stochastic processes: an essay awarded the Adams prize for the University of Cambridge*, Cambridge University Press: Cambridge.

Robert, C. P. & Mengersen, K. L. (1994), Reparameterization issues in mixture modelling and their bearing on the Gibbs sampler, Technical report, Laboratoire de Statistique, CREST, INSEE, Paris.

Saunders, R., Kryscio, R. J. & Funk, G. M. (1979), 'Limiting results for arrays of binary random variables on rectangular lattices under sparseness conditions', *Journal of Applied Probability* **16**, 554–566.

Siegel, S. & Castellan, Jr, N. J. (1988), *Nonparametric statistics for the behavioural sciences*, 2nd edn, McGraw-Hill, New York.

Silverman, B. W., Jones, M. C., Wilson, J. D. & Nychka, D. W. (1990), 'A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography', *J. R. Statist. Soc. Series B* **52**, 271–324.

Simpson, K. & Day, N. (1993), *Field Guide to the Birds of Australia*, fourth edn, Viking, Penguin Books, Australia.

Smith, A. F. M. & Roberts, G. O. (1993), 'Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods', *J. R. Statist. Soc. B* **55**(1), 2–23.

Smyth, G. K., Chakraborty, S., Clark, R. G. & Pettitt, A. N. (1992), 'A stochastic model for Anthracnose development in *stylosanthes scabra*', *Phytopathology* p. 822.

Sokal, A. D. (1989), 'Monte Carlo methods in statistical mechanics: foundations and new algorithms', Lecture Notes, Troisieme cycle de la Physique en Suisse Romande, Department of Physics, New York University.

Stephenson, J. (1964), 'Ising-model spin correlations on the triangular lattice', *J. Math. Phys.* **5**, 1009.

Stiratelli, R., Laird, N. & Ware, J. H. (1984), 'Random effects models for serial observations with binary response', *Biometrics* **40**, 961–71.

Strauss, D. J. (1975), 'Analyzing binary lattice data with the nearest-neighbour property', *J. Appl. Probab.* **12**, 702–712.

Strauss, D. J. (1986), 'On a general class of models for interaction', *SIAM Review* **28**, 513–527.

Swendson, R. H. & Wang, J.-S. (1987), 'Non-universal critical dynamics in Monte Carlo simulations', *Physics Review Letters* **58**, 86–88.

Syozi, I. (1972), *Transformations of Ising models*, Vol. 1. Exact Results of Domb & Green (1972*a*), pp. 269–329.

Temperley, H. N. V. (1972), Two-dimensional Ising models, *in* Domb & Green (1972*a*), pp. 227–267.

The MathWorks, Inc. (1999), *MATLAB Reference Manual*, Natick, MA.

Thompson, C. J. (1988), *Classical Equilibrium Statistical Mechanics*, Oxford Science Publications, Clarendon Press: Oxford.

Tierney, L. (1991), Markov chains for exploring posterior distributions, Technical Report 560, University of Minnesota.

Tierney, L. (1994), 'Markov chains for exploring posterior distributions', *Annals of Statistics* **22**(4), 1701–1762.

Tierney, L. & Kadane, J. B. (1986), 'Accurate approximations for posterior moments and maginal densities', *J. Amer. Statist. Assoc.* **81**(393), 82–86.

Toda, M., Kubo, R. & Saitô (1983), *Statistical Physics I. Equilibrium Statistical Mechanics*, Springer Series in Solid-State Sciences, Springer-Verlag, Hiedelberg.

Torrie, G. M. & Valleau, J. P. (1977), 'Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling', *Journal of Computational Physics* **23**, 187–199.

Upton, G. J. G. & Fingleton, B. (1990), *Spatial data analysis by example*, Wiley series in probability and statistics, reprint of 1985 edition edn, John Wiley & Sons Ltd.

Valleau, J. P. & Card, D. N. (1972), *J. Chem. Phys.* **57**, 5457.

Venema, H. W. (1993), 'Estimation of the parameters of a binary Markov random field on a graph with application to fibre type distributions in a muscle cross-section', *IMA Journal of Mathematics Applied in Medicine and Biology* **10**, 115–133.

Verhagen, A. M. W. (1977), 'A three parameter isotropic distribution of atoms and the hard-core square lattice gas', *J. Chem. Phys.* **67**, 5060–5065.

Vines, S. K., Gilks, W. R. & Wild, P. (1994), Fitting Bayesian multiple random effects models, Technical report, Biostatistics Unit, Medical Research College, Cambridge.

Wakefield, J. (1991), Parameterization issues in Gibbs sampling., *in* 'Workshop on Bayesian Computation via Stochastic Simulation, Columbus, Ohio'.

Walker, P. A. (1990), 'Modelling wildlife distributions using a geographic information system: kangaroos in relation to climate.', *J. Biogeogr.* **17**, 279–289.

Walpole, R. E., Myers, R. H. & Myers, S. L. (1998), *Probability and statistics for engineers and scientists*, 6th edn, Prentice Hall, Upper Saddle River, NJ.

Wang, J.-S. & Swendsen, R. H. (1990), 'Cluster Monte Carlo algorithms', *Phys. A* **167**(3), 565–579.

Ware, J. H. (1985), 'Linear models for the analysis of longitudinal studies', *American Statistician* **39**, 95–101.

Watson, P. G. (1972), Surface and size effects in lattice models, *in* Domb & Green (1972*b*), pp. 101–159.

Weir, I. S. & Pettitt, A. N. (1997), Alternative to the autologistic model using a hidden conditional autoregressive Gaussian process, Technical report, Queensland University of Technology.

Weir, I. S. & Pettitt, A. N. (1999), 'Spatial modelling for binary data using a hidden conditional autoregressive Gaussian process: a multivariate extension of the probit model', *Statistics and Computing* **9**, 77–86.

Welsh, A. H. (1996), 'Robust estimation of smooth regression and spread functions and their derivatives', *Statistica Sinica* **6**, 347–366.

Welsh, A. H., Cunningham, R. B., Donnelly, C. F. & Lindenmayer, D. B. (1996*a*), 'Modelling the abundance of a rare species: statistical models for counts with extra zeroes', *Ecological Modelling* **88**, 297–308.

Welsh, A. H., Cunningham, R. B., Donnelly, C. F. & Lindenmayer, D. B. (1996*b*), 'Modelling the abundance of a rare species; statistical models for counts with extra zeros.', *Ecological Modelling* **88**, 297–308.

West, M. & Harrison, J. (1997), *Bayesian forecasting and dynamic models*, 2nd edn, Springer-Verlag, New York.

West, M., Harrison, J. P. & Migon, H. S. (1985), 'Dynamic generalized linear models and Bayesian forecasting', *J. Amer. Statist. Assoc.* **80**, 73–96.

Whittaker, J. (1990), *Graphical Models in Applied Multivariate Statistics*, John Wiley and Sons, Chichester.

Wilson, K. G. (1976), The renormalization group–introduction, *in* Domb & Green (1976), pp. 1–5.

Winkler, G. (1995), *Image Analysis, Markov Random Fields and Dynamic Monte Carlo Sampling methods: A Mathematical introduction*, Springer Verlag.

Wolff, U. (1989), *Phys. Rev. Lett.* **62**, 361.

Wolpert, R. & Ickstadt, K. (1995), Gamma/Poisson random field models for spatial statistics, Technical report 95–43, Institute of Statistics and Decision Sciences, Duke University, Durham, NC.

Wolpert, R. L. & Ickstadt, K. (1998), 'Poisson/gamma random field models for spatial statistics', *Biometrika* **85**(2), 251–267.

Wortis, M. (1974), Linked cluster expansion, *in* Domb & Green (1974), pp. 114–178.

Wu, H. & Huffer, F. (1997), 'Modelling the distribution of plant species using the Autologistic Regression Model', *Environ. Ecol. Stat.* **4**(1), 49–64.

Yang, C. N. (1952), 'The spontaneous magnetization of a two-dimensional Ising model.', *Physical Review* **87**, 808–816.

Yang, C. N. (1972), Introductory note on phase transitions and critical phenomena, *in* Domb & Green (1972*a*), pp. 1–5.

Yang, C. N. & Lee (1952), 'Statistical theory of equations of state and phase transitions I. Theory of condensation', *Physical Review* **87**, 404–410.

Yu, B. & Mykland, P. (1994), Looking at Markov samplers through CUSUM path plots: a simple diagnostic idea, Technical Report 413, Department of Statistics, University of California, Berkeley.

Zeger, S. L. & Karim, M. R. (1991), 'Generalized linear models with random effects; A Gibbs sampling approach', *J. Amer. Statist. Assoc.* **86**(413), 79–86.

Zeger, S. L. & Qaqish, B. (1988), 'Markov regression models for time series: a quasi-likelihood approach', *Biometrics* **44**, 1019–1031.

Zimmerman, D. L. & Harville, D. A. (1991), 'A random field approach to the analysis of field-plot experiments and other spatial experiments', *Biometrics* **47**, 223–239.